

6€

ESTADÍSTICA ADMINISTRATIVA

Juan Antonio García Ramos, Carmen Ramos González y
Gabriel Ruiz Garzón



UCA

Universidad
de Cádiz

Servicio de Publicaciones



ESTADÍSTICA ADMINISTRATIVA

Juan Antonio García Ramos

Carmen D. Ramos González

Gabriel Ruiz Garzón

ESTADÍSTICA ADMINISTRATIVA

Juan Antonio García Ramos

Carmen D. Ramos González

Gabriel Ruiz Garzón



UCA

Universidad
de Cádiz

Servicio de Publicaciones

© Servicio de Publicaciones. Universidad de Cádiz
Juan Antonio García Ramos, Carmen Ramos Gonzáles, Gabriel Ruiz Garzón

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

<http://www.uca.es/publicaciones>

ISBN13: 978-84-9828-066-1

Depósito Legal: SE-5619-2007 U.E.
Printed by Publigades

1ª edición: 2007
1ª reimpresión: Octubre 2008

Índice general

Prólogo	1
I DISTRIBUCIONES UNIDIMENSIONALES	3
1. Organización de la información	5
1.1. Introducción	5
1.1.1. Breve reseña histórica	5
1.1.2. La Estadística Administrativa: Aplicaciones	6
1.2. Variables estadísticas	7
1.3. Distribuciones de frecuencias. Tipos	8
1.4. Representación numérica	9
1.5. Representaciones gráficas	10
2. Resumen de datos: Medidas de posición	17
2.1. Medidas centrales	17
2.1.1. La media aritmética. Propiedades	17
2.1.2. Otras medias	20
2.1.3. La mediana	22
2.1.4. La moda	24
2.2. Medidas de posición no centrales	26
2.3. Momentos no centrados y centrados	28
3. Resumen de datos: Medidas de dispersión	31
3.1. Medidas de dispersión absoluta	32
3.1.1. Recorridos. Desviaciones medias	33
3.1.2. Varianza y desviación típica	34
3.1.3. Normalización o tipificación	38
3.2. Medidas de dispersión relativa	39
3.2.1. Recorridos	39
3.2.2. El coeficiente de variación de Pearson	39

4. Resumen de datos: Medidas de forma	41
4.1. Medidas de asimetría	41
4.1.1. Coeficiente de asimetría de Fisher	43
4.1.2. Coeficiente de asimetría de Pearson	44
4.2. Medidas de curtosis	46
4.2.1. Introducción al modelo Normal	46
4.2.2. El coeficiente de curtosis de Fisher	47
4.3. Análisis Exploratorio de Datos	49
4.3.1. Diagrama de caja y bigotes	49
4.3.2. Diagrama de tallo y hojas	50
5. Medidas de desigualdad	53
5.1. Medidas de desigualdad o concentración	53
5.2. Estudio gráfico: la curva de Lorenz	54
5.3. Estudio analítico: el índice de Gini	57
Cuestiones, problemas y recursos	59
 II DISTRIBUCIONES BIDIMENSIONALES	 71
6. Variables estadísticas bidimensionales	73
6.1. Distribución conjunta	73
6.2. Distribuciones marginales y condicionadas	76
6.3. Representaciones gráficas	79
6.4. Momentos no centrados y centrados	81
6.5. Independencia de variables estadísticas	83
6.6. Dependencia de variables estadísticas	84
6.7. Dependencia lineal. Covarianza	84
7. Ajustes	89
7.1. Introducción	89
7.1.1. El método de mínimos cuadrados	90
7.2. Ajuste lineal. Función de consumo de Keynes	91
7.3. Ajustes reducibles al caso lineal	94
7.3.1. Ajuste hiperbólico: curvas de demanda	94
7.3.2. Ajuste potencial: función de Cobb-Douglas	95
7.3.3. Ajuste exponencial: modelo de Harrod-Domar	96
7.4. Otros ajustes	98
7.4.1. Exponencial modificada: la Ley de Makeham	98
7.4.2. La curva logística	100

8. Regresión simple	103
8.1. Introducción al concepto de regresión	103
8.2. Regresión de la media	103
8.3. Regresión mínimo-cuadrática	105
8.3.1. Regresión lineal mínimo-cuadrática	106
8.3.2. Propiedades de las rectas de regresión	107
8.4. Línea de Tukey o línea Mediana-Mediana	109
9. Correlación simple	113
9.1. Concepto de correlación	113
9.2. Medidas de correlación	113
9.2.1. Varianza residual. Análisis de los residuos	113
9.2.2. Coeficiente de determinación. Interpretación	114
9.3. Correlación lineal	115
9.3.1. Descomposición de la varianza total	115
9.3.2. El coeficiente de correlación lineal	116
9.4. Bondad de ajuste para otras funciones	117
9.5. Predicción	118
9.6. Correlación espuria	118
Cuestiones, problemas y recursos	119

III NÚMEROS ÍNDICES Y SERIES TEMPORALES 129

10. Números índices	131
10.1. Concepto de número índice	131
10.2. Índices simples	132
10.3. Propiedades de los índices simples	133
10.4. Índices complejos	134
10.4.1. Índices complejos no ponderados	135
10.4.2. Índices complejos ponderados	136
10.5. Índices encadenados	138
11. Índices de precios	139
11.1. Introducción	139
11.2. Índices de precios complejos no ponderados	139
11.3. Índices de precios complejos ponderados	141
11.4. Propiedades de los índices complejos	144
11.5. Enlaces y cambios de base	144
11.6. Deflación de series estadísticas	146
11.7. Variación, repercusión y participación	149

12.Series temporales: análisis descriptivo	153
12.1. Concepto de serie temporal	153
12.2. Descripción de una serie temporal	154
12.3. Análisis de la tendencia	155
12.3.1. Método de ajuste analítico	155
12.3.2. Método de las medias móviles o método mecánico	158
12.4. Análisis de la estacionalidad	160
12.4.1. Método de las medias mensuales o método analítico	160
12.4.2. Método de las medias móviles o método mecánico	163
Cuestiones, problemas y recursos	169

IV PROBABILIDAD **181**

13.Probabilidad	183
13.1. Experimentos aleatorios. Definiciones	183
13.2. Álgebra de sucesos. Propiedades	184
13.3. Diversas concepciones de probabilidad	187
13.4. Propiedades	189
14.Probabilidad condicionada	191
14.1. Probabilidad condicionada. Propiedades	191
14.2. Teorema del producto	193
14.3. Sucesos dependientes e independientes	193
14.4. Teorema de la probabilidad total	195
14.5. Teorema de Bayes	196
Cuestiones, problemas y recursos	199

V MODELOS DE DISTRIBUCIONES **205**

15.Variables aleatorias	207
15.1. Variable aleatoria: concepto	207
15.2. Función de distribución. Propiedades	208
15.3. Variables aleatorias discretas	209
15.4. Variables aleatorias continuas	210
15.5. Características de las variables aleatorias	213
15.5.1. Introducción	213
15.5.2. Esperanza matemática. Propiedades	213
15.5.3. Momentos	215
15.5.4. Varianza y desviación típica. Propiedades	216

16. Modelos probabilísticos discretos	219
16.1. Introducción	219
16.2. La distribución Binomial	220
16.3. La distribución de Poisson	222
16.3.1. Aproximación de Poisson a la distribución Binomial	224
17. La distribución Normal	227
17.1. Introducción	227
17.2. Definición y propiedades	228
17.3. Distribución Normal tipificada	231
17.4. Uso de tablas	231
17.5. Teorema Central del Límite	233
17.6. Aproximaciones mediante la Normal	234
Cuestiones, problemas y recursos	239
 VI MÉTODOS DE INFERENCIA ESTADÍSTICA	 251
18. Introducción al Muestreo	253
18.1. Definiciones	253
18.2. Introducción a la Teoría de Muestras	256
18.3. Muestreos no probabilísticos	257
18.4. Muestreos probabilísticos	258
18.4.1. Muestreo aleatorio simple	258
18.4.2. Muestreo aleatorio con reemplazamiento	259
18.4.3. Muestreo estratificado	259
18.4.4. Muestreo por conglomerados unietápico	260
18.4.5. Muestreo de conglomerados con submuestreo	261
18.4.6. Muestreo sistemático	262
18.4.7. Muestreo bifásico	263
18.5. Otros tipos de muestreo	264
18.6. Métodos muestrales en el tiempo	264
19. La Administración y las Estadísticas	267
19.1. El sistema estadístico en las Administraciones	267
19.2. Las estadísticas demográficas	268
19.2.1. Censos y Padrón	268
19.2.2. Otras encuestas demográficas	269
19.3. Las estadísticas económicas	269
19.3.1. La Encuesta de Población Activa	269
19.3.2. La Encuesta de Presupuestos Familiares	270
19.3.3. El nuevo Índice de Precios de Consumo	271
19.3.4. El Índice de Consumo Armonizado	274

19.3.5. El Índice de Producción Industrial	275
19.3.6. El Índice de Precios Industriales	276
19.3.7. Otras encuestas económicas	277
19.4. Las estadísticas sociales	278
19.4.1. El Panel de Hogares	278
19.4.2. Encuestas Turísticas	279
19.5. Otras encuestas públicas	280
20. Muestreo en poblaciones normales	283
20.1. La distribución χ^2 de Pearson	283
20.1.1. Distribución de la varianza muestral	285
20.2. Distribución t de Student	285
20.2.1. Distribución del estadístico media muestral	287
20.2.2. Distribución de la diferencia de medias muestrales	288
20.3. Distribución F de Fisher-Snedecor	288
20.3.1. Distribución del cociente de varianzas muestrales	290
21. Estimación	291
21.1. Estimación puntual paramétrica	291
21.1.1. El método analógico	291
21.1.2. El método de los momentos	292
21.2. Estimación por intervalos de confianza	292
21.2.1. Concepto de intervalo de confianza	292
21.2.2. Método del pivote	294
21.3. Intervalo para la media	295
21.3.1. Varianza conocida	295
21.3.2. Varianza desconocida y muestra pequeña	297
21.3.3. Varianza desconocida y muestra grande	298
21.4. Intervalo de confianza para la varianza	299
21.5. Intervalo para las medias	300
21.5.1. Varianzas conocidas	300
21.5.2. Varianzas desconocidas, iguales y muestras pequeñas	301
21.5.3. Varianzas desconocidas, distintas y muestras pequeñas	301
21.5.4. Varianzas desconocidas y muestras grandes	301
21.6. Intervalo para las medias de datos apareados	303
21.7. Intervalo para la razón de varianzas	304
21.8. Intervalos de confianza asintóticos	305
21.8.1. Intervalo de confianza para la proporción	305
21.8.2. Intervalo para la diferencia de proporciones	308

22. Contrastes de hipótesis	311
22.1. Introducción	311
22.2. Pasos para la realización de un contraste	314
22.3. Relación entre intervalos y contrastes	318
22.4. Contrastes para la media	318
22.4.1. Varianza conocida	318
22.4.2. Varianza desconocida y muestra pequeña	320
22.4.3. Varianza desconocida y muestra grande	321
22.5. Contraste para la varianza	322
22.6. Contrastes para dos medias	323
22.6.1. Varianzas conocidas	323
22.6.2. Varianzas desconocidas, iguales y $n_1 < 30$ ó $n_2 < 30$	323
22.6.3. Varianzas desconocidas, distintas y $n_1 < 30$ ó $n_2 < 30$	323
22.6.4. Varianzas desconocidas y $n_1 \geq 30$ ó $n_2 \geq 30$	324
22.7. Contraste para 2 medias. Datos apareados	325
22.8. Contraste para las varianzas	326
22.9. Contrastes asintóticos	328
22.9.1. Contraste para la proporción	328
22.9.2. Contraste para la igualdad de proporciones	329
Cuestiones, problemas y recursos	331
Bibliografía	341

Prólogo

Presentamos este manual fruto de nuestra labor pedagógica en la Diplomatura de Gestión y Administración y Pública, y con el deseo fundamental de que resulte útil a, no sólo a nuestros estudiantes, sino a cualquier estudiante de Primer Ciclo universitario, donde un curso básico de Estadística tenga cabida. Este volumen que el lector tiene en sus manos, viene a cubrir la práctica inexistencia de manuales dedicados a la Estadística Administrativa como tal, debido entre otros factores, a la singularidad que la Diplomatura de Gestión y Administración Pública tiene en la Universidad Española.

No obstante, la relación entre la Administración y la Estadística ha sido muy intensa. No olvidemos que los términos Estadística y Estado comparten la misma raíz latina o que la Estadística fue llamada en sus comienzos Política Aritmética.

Esta obra se ha estructurado en seis partes, finalizando cada una de ellas con una colección de problemas y cuestiones propuestas, que proporcionan la oportunidad de repetir los principios básicos de la asignatura, elemento fundamental para un aprendizaje eficaz de la misma, así como algunas direcciones electrónicas dónde encontrar interesantes recursos.

Las primeras partes de este ejemplar están dedicadas a la descripción de variables con modelos frecuencialistas, los números índices y las series temporales, conformando lo que se entiende por Estadística Descriptiva. Se trata de la parcela más sencilla de la Estadística y, quizás por eso, es a menudo infravalorada. En todo caso, la descripción constituye la fase previa de cualquier análisis estadístico y su dominio ha de condicionar el éxito de los resultados finales: precisión de los pronósticos, veracidad de conclusiones, etc. Destacaremos de estas partes los apartados dedicados al Análisis Exploratorio de Datos, al análisis de la desigualdad y los ajustes a la curva logística y ley de Makeham, esto último de tanta importancia en el estudio de las poblaciones humanas.

Muchas de las observaciones realizadas por un científico contienen elementos procedentes del azar, que impide llegar a la certeza absoluta y, por tanto, el azar y su medición, a través de la probabilidad, serán objeto de estudio en la cuarta parte de este manual.

La quinta parte la dedicaremos a la construcción de modelos de distribuciones de probabilidad que pudieran representar el comportamiento de diferentes fenómenos aleatorios que aparecen en el mundo real.

Dedicaremos la última parte al estudio de la Inferencia Estadística, comprendiendo técnicas como la estimación y el contraste de hipótesis, con objeto de proporcionar al investigador instrumentos para la toma de decisiones cuando prevalece el azar. También dedicaremos un capítulo a la introducción al muestreo en poblaciones finitas y otro al estudio de las características de las principales encuestas que realiza la Administración.

Somos conscientes de la dificultad que entraña para algunos de nuestros estudiantes el tratamiento numérico, inevitable, que el estudio de la Estadística comporta. Hemos intentado hacerlo lo más asequible posible pero manteniendo un mínimo de rigor.

Es un deber expresar nuestro agradecimiento a todos aquellos que nos han sugerido modificaciones o añadidos o que simplemente nos han animado en la realización de este manual: alumnos, compañeros entrañables como la profesora Antonia Castaño Martínez, amigos y familiares. A todos vosotros, gracias.

Jerez de la Frontera, Abril 2006, Los Autores.

Parte I

DISTRIBUCIONES UNIDIMENSIONALES

Capítulo 1

Organización de la información

1.1. Introducción

Definición 1.1 *La Estadística Descriptiva trata de organizar, representar y resumir un conjunto de datos de manera que pueda ser extraída la máxima información procedente de ellos.*

1.1.1. Breve reseña histórica

Desde la antigüedad, la Administración ha recogido información sobre la población y la riqueza que existía en sus dominios. Entre los trabajos censales más antiguos, podemos citar:

1. En Sumeria, entre los 5000 y 2000 a. de J.C., se inscribían en tablas de arcilla, con caracteres cuneiformes, la relación de hombres y sus pertenencias.
2. En Egipto, por ejemplo, bajo el reinado de Amasis II, cada persona debía declarar su profesión y su fuente de ingresos.
3. En Roma, el primer censo fue realizado a instancias de Servio Tulio (578-534 a. de J.C.), para clasificar a los ciudadanos en función de sus ingresos para poder girar después los correspondientes impuestos, siendo muy conocido el Censo de Augusto en el inicio de la era cristiana. En el Imperio Romano, la preocupación por las actividades de los recuentos de los individuos y bienes de la Administración, tenía una clara intención tributaria y/o militar. No olvidemos que la palabra censo proviene de la palabra latina *censere*, que significa gravar con un impuesto.

4. En la Edad Media, el contador Alonso de Quintanilla realizó, por encargo de Isabel La Católica, una copiosa recogida de datos acerca de las riquezas y la población entre 1477 y 1479.
5. El primer censo que se realizó en América fue realizado en Perú en 1548 bajo la dirección del virrey D. Pedro de la Gasca.
6. Más recientemente, en España son famosos los censos de la Ensenada de 1749 y el de Floridablanca de 1787.

La escuela alemana, desarrollada a lo largo de los siglos XVII y XVIII, pone énfasis especial en la Estadística como la descripción comparativa de los Estados.

Paralelamente a la escuela alemana, surge la escuela inglesa, encabezada por Graunt y Petty. Su objetivo era el conocimiento demográfico de la población londinense, que había ido disminuyendo paulatinamente por efecto de las sucesivas plagas de peste que asolaron la ciudad.

Cerca de las escuelas alemana e inglesa, la escuela belga encabezada por Quetelet (1796-1874), quién recolectó toda clase de datos sociales y describió su distribución de frecuencias en términos de la ley normal e introdujo el concepto de *hombre medio*.

A partir del siglo XVII, y con total independencia del análisis que acabamos de describir, se inician los estudios sobre la Teoría de Probabilidades ligado a los juegos de azar. En este tema son importantes los trabajos de Pascal, Fermat, De Moivre y Huygens entre otros.

A lo largo del siglo XIX tenemos a Laplace, Gauss, Poisson y a los componentes de la Escuela de San Petersburgo, con importantes aportaciones a la demostración del Teorema Central del Límite.

A finales del siglo XIX y comienzos del XX, comienza una interacción entre el Cálculo de Probabilidades y la Estadística. Pearson, padre e hijo, Fisher y Newmann, contribuyeron decisivamente a la llamada Estadística Inferencial.

A partir de 1950 comienza la época moderna de la Estadística con la aparición de los ordenadores e Internet. Un gran paso en la visión integradora de la Estadística Descriptiva e Inferencial fue dado por Tukey y Mosteller al desarrollar el Análisis Exploratorio de Datos. La filosofía es entender los especiales rasgos de los datos antes de efectuar ningún procedimiento estadístico.

1.1.2. La Estadística Administrativa: Aplicaciones

Seguidamente, y sin ánimo de ser exhaustivo, reseñaremos algunas utilidades de la Estadística, dentro de la Administración. Ya hemos visto en el anterior apartado, que desde la Antigüedad, la Administración ha utilizado la Estadística para el buen gobierno de sus estados.

Más recientemente, a modo de ejemplo, en materia impositiva, en 1986, el Congreso de los Estados Unidos convirtió en permanente una subida del cien por cien del impuesto del tabaco. El principal motivo fue la necesidad de ingresos

adicionales. Una segunda consideración fue la creencia, basada en un Análisis de Regresión, que una bajada del impuesto aumentaría el número de adictos al tabaco y, a la larga, un mayor número de enfermos y muertes.

Otro tema que preocupa y ocupa a la Administración fue la Salud Pública. Continuamente se están realizando experimentos para evaluar la efectividad de los distintos medicamentos. En este sentido citaremos uno de los mayores experimentos médicos y estadísticos, que involucró a más de un millón de niños, en el año 1954 en los Estados Unidos, para evaluar la bondad de la vacuna de Salk contra la poliomielitis. Estadísticamente se pudo aconsejar la vacunación de toda la población estadounidense.

De igual manera la Estadística agrupa en su seno diversos métodos estadísticos para el análisis de la desigualdad en el reparto de la renta total de un país. Así, Conrado Gini obtuvo su índice de concentración en relación con los ingresos y Lorenz propuso un método gráfico para medir la concentración de los ingresos.

De indudable interés para la Administración Pública es también la medida de la inflación a través de la Teoría de los Números Índices. Estadísticos como E. Laspeyres o F.Y. Edgeworth dieron nombres a famosos índices. Irving Fisher propuso la búsqueda del número índice ideal. A los índices de precios se unieron después los índices de producción, debidos a Burns y los índices de salarios y empleo, de Bowley y Wood.

Es objetivo fundamental de la Administración el prolongar los ciclos de bonanza y prever los cambios de tendencia de nuestra economía. Los indicadores cíclicos son series o conjunto de datos ordenados temporalmente, como la demanda eléctrica, el consumo de fuel, la entrada de extranjeros o la matriculación de automóviles, etc., de alta sensibilidad cíclica, cuya utilidad es medir y prever las fluctuaciones económicas y los puntos de cambio de tendencia del ciclo económico.

El análisis de los indicadores económicos realizado por la Administración, ha demostrado la factibilidad del procedimiento y el valor potencial para observar y apreciar fluctuaciones internacionales en las tasas de crecimiento económico y las tendencias asociadas a los niveles de precios, inversión de capital y empleo.

1.2. Variables estadísticas

Definición 1.2 *Las características que poseen los elementos de una población y que van a ser objeto de estudio estadístico reciben el nombre de variables estadísticas.*

Tipos de variables

- (a) *Cualitativas o atributos*: son aquellas características no expresables numéricamente. Sus posibles valores se llaman modalidades o categorías. No se

pueden asociar naturalmente a un número y no se pueden hacer operaciones algebraicas con ellos.

EJEMPLO 1.1 *Sexo, nivel de estudios, etc.*

(b) *Cuantitativas*: son aquellas características expresables numéricamente.

- (i) Variables cuantitativas discretas. Son aquellas variables cuyos posibles valores constituyen un conjunto de cardinal finito o a lo sumo infinito numerable.

EJEMPLO 1.2 *Tamaño de una familia, número de obreros, etc.*

- (ii) Variables cuantitativas continuas. Son aquellas que pueden tomar los infinitos valores de un intervalo, es decir, si entre dos valores son posibles infinitos valores intermedios.

EJEMPLO 1.3 *Ingresos mensuales, gasto en vestido, etc.*

OBSERVACIÓN 1.1 *Las variables suelen representarse con las últimas letras del alfabeto y escritas en mayúsculas $X, Y, Z, \dots$*

1.3. Distribuciones de frecuencias. Tipos

A partir de un conjunto de datos queremos clasificarlos de modo que la información contenida en ellos quede presentada de forma clara, concisa y ordenada. A lo largo del presente tema nos ocuparemos de esto. Así definiremos:

Definición 1.3 *Frecuencia ordinaria o absoluta del valor x_i de la variable es el número de veces que se presenta dicho valor en el conjunto de datos. Se representa por n_i .*

Definición 1.4 *Frecuencia absoluta acumulada del valor x_i de la variable es el número de datos que hay iguales o inferiores al considerado. Se representa por N_i .*

Definición 1.5 *Frecuencia relativa, $f_i = \frac{n_i}{N}$ donde $N = \sum_{i=1}^k n_i$ es la frecuencia total, o número total de datos.*

Definición 1.6 *Frecuencia relativa acumulada, $F_i = \frac{N_i}{N}$.*

Definición 1.7 *Distribución de frecuencias es el conjunto de los valores que presenta una variable estadística junto con sus frecuencias. En general, escribiremos $\{(x_i; n_i)\}_{i=1,2,\dots,k}$.*

Tipos de distribuciones de frecuencias:

- (a) Distribuciones de tipo discreto o de datos sin agrupar.
- (b) Distribuciones de tipo continuo o de datos agrupados.

1.4. Representación numérica

Los datos se representan numéricamente en forma de tabla, colocando los diferentes valores de la variable, x_i , acompañados de sus frecuencias.

- (a) Datos sin agrupar y cualitativos:

x_i	n_i
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_k	n_k

EJEMPLO 1.4 Se considera la variable "número de miembros", estudiada a 20 familias:

2 ; 3 ; 4 ; 5 ; 8 ; 2 ; 4 ; 5 ; 7 ; 2 ; 1 ; 3 ; 6 ; 4 ; 2 ; 5 ; 5 ; 1 ; 3 ; 4

x_i	n_i	N_i	f_i	F_i
1	2	2	0,10	0,10
2	4	6	0,20	0,30
3	3	9	0,15	0,45
4	4	13	0,20	0,65
5	4	17	0,20	0,85
6	1	18	0,05	0,90
7	1	19	0,05	0,95
8	1	$\sum_{i=1}^k n_i = N = 20$	0,05	$\sum_{i=1}^k f_i = 1$

- (b) Datos agrupados o de tipo continuo. Las variables cuantitativas continuas o las discretas que presenten un gran número de valores distintos, para mayor comodidad en el tratamiento de la información, se agruparán en clases o intervalos. Para ello tendremos en cuenta lo siguiente:

- (i) Dividiremos el recorrido de la variable, $R = \max_i x_i - \min_i x_i$, en clases o intervalos que no se solapen.

- (ii) Consideraremos intervalos abiertos a la izquierda y cerrados a la derecha, $(l_{i-1}, l_i]$, para evitar que algún x_i pueda pertenecer a más de un intervalo.
- (iii) Se define la amplitud del intervalo $(l_{i-1}, l_i]$ como $c_i = l_i - l_{i-1}$.
- (iv) Como representante de la clase, o valor ideal de la misma, se toma la *marca de clase*, que es el punto medio del intervalo: $x_i = \frac{l_{i-1} + l_i}{2}$.

Las distribuciones agrupadas en intervalos se resumirán en una tabla con el formato siguiente:

$l_{i-1} - l_i$	n_i	x_i	c_i
$l_0 - l_1$	n_1	$x_1 = \frac{l_0 + l_1}{2}$	$c_1 = l_1 - l_0$
$l_1 - l_2$	n_2	$x_2 = \frac{l_1 + l_2}{2}$	$c_2 = l_2 - l_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$l_{k-1} - l_k$	n_k	$x_k = \frac{l_{k-1} + l_k}{2}$	$c_k = l_k - l_{k-1}$

OBSERVACIÓN 1.2

- (a) *El agrupamiento de los datos da lugar a cierta pérdida de información pero con ello se gana en manejabilidad de los mismos.*
- (b) *El número de intervalos y las amplitudes de los mismos deben ser escogidos convenientemente. En la práctica, es frecuente la elección de intervalos de amplitud constante, ya que con ello se facilita el cálculo de la mayoría de las características descriptivas que analiza la Estadística.*

1.5. Representaciones gráficas

Su finalidad consiste en presentar, a golpe de vista, el comportamiento de la distribución. Se usan, por tanto, como complemento del trabajo estadístico, y a veces, como punto de partida para el posterior análisis estadístico.

Tipos de gráficos:

- (a) Para variables cualitativas preferentemente: basan su construcción en establecer proporcionalidad entre áreas y frecuencias.

EJEMPLO 1.5 *Se considera el estudio del atributo "lugar de procedencia de los turistas llegados a la provincia de Cádiz durante el año 1997".*

<i>Lugar de procedencia</i>	<i>Número de turistas (n_i)</i>
<i>Andalucía</i>	659 707
<i>Resto de España</i>	914 990
<i>Unión Europea</i>	845 123
<i>Resto del Mundo</i>	161 232
	$N = 2\,581\,052$

FUENTE: IEA. Encuesta de Coyuntura turística (ECTA)

- (i) *Diagrama de sectores o de pastel*: El área de cada sector circular que representa a cada categoría del atributo debe ser proporcional a su frecuencia absoluta.

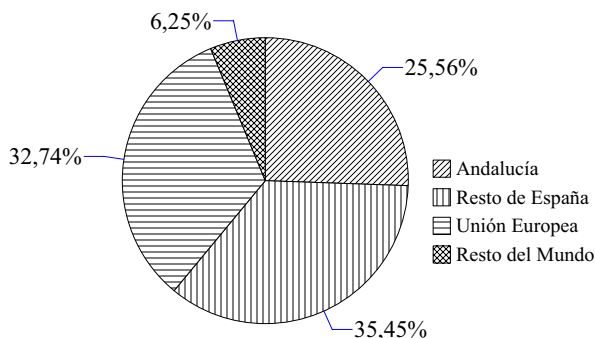


Figura 1.1: Diagrama de sectores

- (ii) *Diagramas de rectángulos*: Se construyen dibujando sobre cada categoría del atributo rectángulos de igual base y altura la frecuencia absoluta (o relativa) de la misma. Por tanto su superficie será también proporcional a la frecuencia absoluta (o relativa) de la categoría correspondiente.
- (iii) *Diagrama de Pareto*: Es un diagrama de rectángulos donde se ordenan las categorías del atributo de mayor a menor frecuencia. Posteriormente se procede a acumular las frecuencias sobre cada categoría trazando una línea que nace del primer rectángulo y que expresa en que medida contribuye cada categoría a la frecuencia total. El diagrama de Pareto nos permite identificar las modalidades "vitales" del análisis.

(b) Para variables cuantitativas:

- (i) Para datos sin agrupar. Se realizan mediante un sistema de ejes cartesianos representando en el eje de abscisas los valores de la variable y

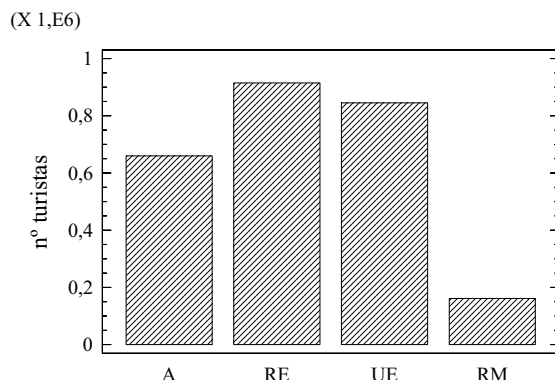


Figura 1.2: Diagrama de rectángulos

en el de ordenadas las frecuencias correspondientes. El más conocido es el *diagrama de barras*. Sobre el eje de ordenadas se representan las frecuencias (absolutas o relativas).

EJEMPLO 1.6 Se considera el estudio de la variable $X = \text{"número de componentes por familia"}$ realizado sobre 75 familias de Jerez

x_i	1	2	3	4	5	6	7	9
n_i	6	11	11	20	15	8	3	1

- (ii) Para datos agrupados. Representaremos en el eje de abscisas los intervalos en los que se agrupan los valores de la variable. Sobre cada uno de ellos dibujaremos un rectángulo de área proporcional a la frecuencia considerada.

- (1) *Histograma de frecuencias*. En este caso se toman como alturas de los rectángulos los valores $h_i = \frac{f_i}{c_i}$.

OBSERVACIÓN 1.3

- (a) Es importante hacer notar que, con las alturas que hemos tomado, el área total del histograma es uno. Este aspecto nos será de gran interés para comprender conceptos posteriores.
- (b) La mayoría de los autores consideran como alturas los valores n_i , para intervalos de amplitud constante o el cociente $n_i/c_i = d_i$ (**densidad de frecuencias**), para intervalos de amplitud variable.
- (c) La decisión sobre el número de intervalos que debe tomarse para construir el histograma es decisiva para la comprensión del mismo.

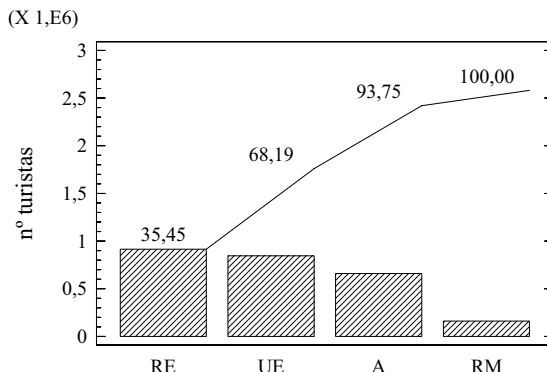


Figura 1.3: Diagrama de Pareto

- (d) Tomando como base el histograma y para obtener una visión de la forma que puede tener la distribución de frecuencias de la variable cuando el número de observaciones fuera muy grande, se emplea el llamado **polígono de frecuencias**. Se forma uniendo mediante segmentos los puntos medios de las bases superiores de los rectángulos del correspondiente histograma. Tiene la propiedad de que el área que encierran coincide con la del histograma que los sustenta, siempre que los intervalos sean todos de igual amplitud.

EJEMPLO 1.7 Consideremos el estudio de la variable $X = \text{"Gasto total en miles de pesetas"}$ de 50 familias

$l_{i-1} - l_i (\times 100)$	n_i	h_i
0 - 50	1	0,0004
50 - 100	8	0,0032
100 - 150	7	0,0028
150 - 200	11	0,0044
200 - 250	11	0,0044
250 - 300	8	0,0032
300 - 350	3	0,0012
350 - 400	1	0,0004

A continuación representamos su histograma y su polígono de frecuencias.

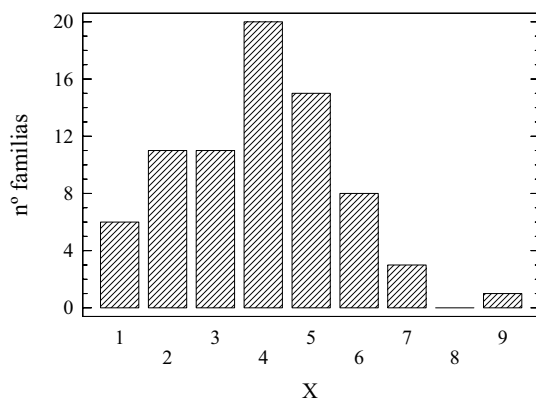
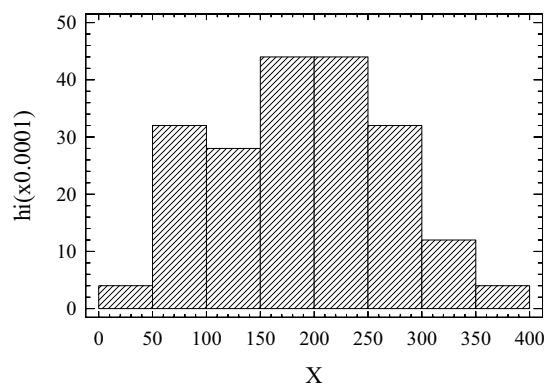
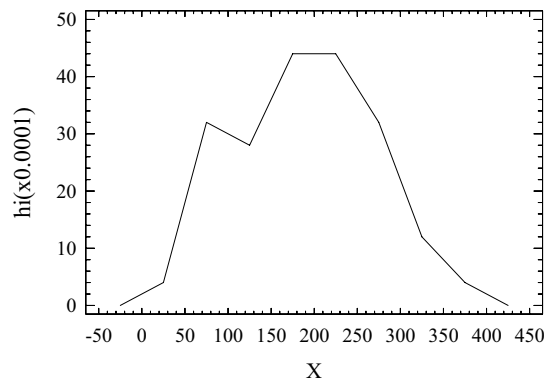


Figura 1.4: Diagrama de barras





Capítulo 2

Resumen de datos: Medidas de posición

El presente tema se ocupa de la etapa del análisis estadístico que consistía en la determinación de medidas o parámetros que intenten resumir la cantidad de información. Se trataría de dar una idea global de cómo es la distribución sin tener que recordar todos los datos.

Las medidas de posición son coeficientes, de tipo promedio o no, que tratan de representar a la distribución de partida.

2.1. Medidas centrales

Definición 2.1 *Las medidas de posición central sirven para representar globalmente el comportamiento de los datos observados y localizar la distribución de frecuencias.*

2.1.1. La media aritmética. Propiedades

Definición 2.2 *La media aritmética de la distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$ es el valor:*

$$\bar{x} = \frac{\sum_{i=1}^k x_i n_i}{N} = \sum_{i=1}^k x_i f_i$$

OBSERVACIÓN 2.1 *Si la variable viene agrupada en intervalos, la media aritmética se calcula utilizando las marcas de clase. El valor resultante de la media aritmética vendrá influenciado por la elección de los intervalos, siendo mayor la precisión cuanto menores sean las longitudes de los intervalos.*

Propiedades de la media aritmética

- (a) La suma de las desviaciones de los valores de la variable respecto a su media es cero:

$$\sum_{i=1}^k (x_i - \bar{x})n_i = \sum_{i=1}^k x_i n_i - N\bar{x} = N\bar{x} - N\bar{x} = 0$$

De esta manera podemos considerar a la media como el *centro de gravedad* de la distribución.

- (b) Teorema de König: La media de las desviaciones al cuadrado de los valores de la variable respecto a una constante p , cualquiera, se hace mínima cuando $p = \bar{x}$. Vamos a comprobarlo:

$$\begin{aligned} D(p) &= \sum_{i=1}^k (x_i - p)^2 \cdot \frac{n_i}{N} = \\ &= \sum_{i=1}^k (x_i - \bar{x} + \bar{x} - p)^2 \cdot \frac{n_i}{N} = \sum_{i=1}^k [(x_i - \bar{x}) - (p - \bar{x})]^2 \cdot \frac{n_i}{N} = \\ &= \sum_{i=1}^k (x_i - \bar{x})^2 \cdot \frac{n_i}{N} + (p - \bar{x})^2 \cdot \frac{\sum_{i=1}^k n_i}{N} - 2(p - \bar{x}) \cdot \sum_{i=1}^k (x_i - \bar{x}) \cdot \frac{n_i}{N} = \\ &= \sum_{i=1}^k (x_i - \bar{x})^2 \cdot \frac{n_i}{N} + (p - \bar{x})^2 \end{aligned}$$

El valor de p que hace mínima esta expresión es $p = \bar{x}$, ya que en este caso el segundo sumando se anula y el primero no depende del valor de p .

- (c) Si de un conjunto de valores obtenemos dos o más subconjuntos disjuntos, la media aritmética de todo el conjunto se relaciona con las medias aritméticas de los diferentes subconjuntos de la forma siguiente:

$$\begin{array}{cc} \hline x_i & n_i \\ \hline x_1 & n_1 \\ \vdots & \vdots \\ x_h & n_h \\ \hline & N_{(1)} \\ \\ x_{h+1} & n_{h+1} \\ \vdots & \vdots \\ x_k & n_k \\ \hline & N_{(2)} \end{array}$$

$$N = N_{(1)} + N_{(2)}$$

$$\bar{x} = \frac{\sum_{i=1}^k x_i n_i}{N} = \frac{\sum_{i=1}^h x_i n_i + \sum_{i=h+1}^k x_i n_i}{N} = \frac{N_{(1)}\bar{x}_1 + N_{(2)}\bar{x}_2}{N}$$

Ventajas de la media aritmética

- (a) Considera todos los valores para su cálculo.
- (b) De existir, es única.
- (c) No se ve afectada por el orden en que vengan dados los datos.
- (d) Es el centro de gravedad de toda la distribución, es decir, representa a todo el conjunto de valores observados.

Inconvenientes de la media aritmética

- (a) Es calculable, salvo que la distribución carezca del extremo inferior del primer intervalo o del extremo superior del último.
- (b) No es robusta, es decir, valores de la variable anormalmente extremados, también llamados valores atípicos u *outliers*, pueden distorsionar la media aritmética. La solución será eliminarlos o calcular otra medida de posición que no se vea afectada por los mismos.

EJEMPLO 2.1 *Consideremos la distribución siguiente y veamos el efecto ejercido por el outlier*

x_i	1	2	3	100
n_i	2	5	2	1

$$\bar{x} = \frac{2 + 10 + 6 + 100}{10} = 11,8$$

Efecto sobre la media aritmética de una transformación lineal

Nos disponemos a estudiar cómo se ve afectada la media ante una transformación lineal que puedan sufrir los datos.

- (a) Dada una distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$, vamos a considerar una nueva distribución $\{(x'_i; n_i)\}_{i=1,2,\dots,k}$, donde $x'_i = x_i + b$ para todo valor de i . Se verifica que:

$$\bar{x'} = \frac{\sum_{i=1}^k x'_i n_i}{N} = \frac{\sum_{i=1}^k (x_i + b) n_i}{N} = \frac{\sum_{i=1}^k x_i n_i}{N} + b \cdot \frac{\sum_{i=1}^k n_i}{N} = \bar{x} + b$$

Por tanto, si a todos los valores de una variable le sumamos una constante b , la media aritmética queda también aumentada en esa constante. Es decir, *la media aritmética queda afectada por los cambios de origen*.

- (b) Dada una distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$, vamos a considerar una nueva distribución $\{(x'_i; n_i)\}_{i=1,2,\dots,k}$, donde $x'_i = ax_i$ para todo valor de i . Se verifica que:

$$\overline{x'} = \frac{\sum_{i=1}^k x'_i n_i}{N} = \frac{\sum_{i=1}^k (ax_i) n_i}{N} = a \cdot \frac{\sum_{i=1}^k x_i n_i}{N} = a \cdot \bar{x}$$

Por tanto, si todos los valores de una variable los multiplicamos por una constante a , la media aritmética queda también multiplicada por esa constante. Es decir, *la media aritmética queda afectada por los cambios de escala*.

Consecuencia

Dada una distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$, si consideramos una nueva distribución $\{(x'_i; n_i)\}_{i=1,2,\dots,k}$ que sea una transformación lineal de la primera, es decir, $x'_i = ax_i + b$ para todo valor de i , entonces se verifica que:

$$\overline{x'} = a\bar{x} + b$$

EJEMPLO 2.2 Si tenemos

$$\left. \begin{array}{l} \bar{x} = 10 \\ x'_i = 3x_i + 2 \end{array} \right\} \Rightarrow \overline{x'} = 3\bar{x} + 2 = 3 \times 10 = 32$$

2.1.2. Otras medias

La media geométrica

La media geométrica se emplea cuando los valores de la variable son no negativos, su número es pequeño y las variaciones entre ellos son grandes, como ocurre en determinadas situaciones económicas (precios, patrimonios de un conjunto de familias, etc.), para promediar porcentajes, tasas, etc. Es decir, en aquellos casos en los que la variable representa variaciones acumulativas, así como cuando dichos valores se encuentran en progresión geométrica.

Definición 2.3 Se define la media geométrica de la distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$, como la expresión:

$$G = \sqrt[N]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}} = \left[\prod_{i=1}^k x_i^{n_i} \right]^{1/N}$$

Ventajas de la media geométrica

- (a) Si existe es única.
- (b) En su cálculo intervienen todos los valores de la distribución.
- (c) Es menos sensible que la $\bar{x}$ a los valores extremos, por su carácter de producto.

EJEMPLO 2.3 *Sobre los datos del ejemplo 2.1*

$$G = \sqrt[10]{1^2 \cdot 2^5 \cdot 3^2 \cdot 100^1} = 2,7921$$

Inconvenientes de la media geométrica

- (a) Su significado es menos intuitivo que el de la $\bar{x}$.
- (b) Cómputo difícil.
- (c) Un sólo dato nulo, hace que el valor de G sea cero, y puede dejar de ser representativa.

La media armónica

Hay ocasiones en que los valores de una variable vienen expresados en términos de los de otra que es inversamente proporcional o recíproca de la primera (precio y poder adquisitivo, velocidad y tiempo, demanda y precio de cierto producto, etc.). En este caso necesitamos un promedio que tenga en cuenta esta relación de reciprocidad.

Definición 2.4 *Dada la distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$, la media armónica H viene dada por la expresión siguiente:*

$$H = \frac{N}{\sum_{i=1}^k \frac{1}{x_i} \cdot n_i}$$

Ventajas de la media armónica

- (a) Si existe es única.
- (b) En su cálculo intervienen todos los valores de la distribución.
- (c) Se ve afectada menos que la $\bar{x}$ por los valores extremadamente elevados.

EJEMPLO 2.4 *Sobre los datos del ejemplo 2.1*

$$H = \frac{10}{\frac{2}{1} + \frac{5}{2} + \frac{2}{3} + \frac{1}{100}} = 1,9317$$

Inconvenientes de la media armónica

- (a) Se ve afectada por los valores pequeños.
- (b) Si algún valor $x_i = 0$, no se puede calcular.

OBSERVACIÓN 2.2 *Se puede demostrar que para una misma distribución de frecuencias de valores positivos se verifica que:*

$$H \leq G \leq \bar{x}$$

La media recortada

La noción de media recortada es uno de los posibles remedios para la falta de robustez de la media. Consiste en moderar el efecto de los outliers sobre la media aritmética suprimiendo un tanto por ciento de las observaciones más extremas.

Definición 2.5 *La media recortada al α por ciento R_α es la media de los datos que quedan después de eliminar el α por ciento de los datos más grandes y el α por ciento de los más pequeños.*

EJEMPLO 2.5 *Sobre los datos del ejemplo 2.1 si calculo la media recortada al 10 por ciento, como tenemos 10 datos suprimo el menor y el mayor, quedándome la siguiente distribución*

x_i	1	2	3
n_i	1	5	2

$$R_{10} = \frac{1 + 10 + 6}{8} = 2,125$$

OBSERVACIÓN 2.3 *Para que verdaderamente sea eficaz esta media recortada al α por ciento, la proporción de outliers en cada extremo debe ser menor que ese α .*

2.1.3. La mediana

Definición 2.6 *La mediana, M_e , es un valor tal que, ordenados los valores de la distribución de menor a mayor, separa a los mismos en dos partes que contienen el mismo número de datos.*

Cálculo de la mediana

- (a) En distribuciones de frecuencias sin agrupar.

$$(i) \text{ Si } N_{i-1} < \frac{N}{2} < N_i \Rightarrow M_e = x_i$$

$$(ii) \text{ Si } N_i = \frac{N}{2} \Rightarrow M_e = \frac{x_i + x_{i+1}}{2}$$

EJEMPLO 2.6 Consideremos los datos siguientes sobre el tamaño de un conjunto de 75 y 76 familias, respectivamente:

x_i	n_i	N_i
1	6	6
2	11	17
3	11	28
<u>4</u>	20	48
5	15	63
6	8	71
7	3	74
9	1	75

$$N/2 = 37,5$$

$$M_e = 4$$

x_i	n_i	N_i
1	5	5
2	10	15
3	23	38
<u>4</u>	15	53
5	11	64
6	6	70
7	4	74
9	2	76

$$N/2 = 38$$

$$M_e = \frac{3+4}{2} = 3,5$$

(b) En las distribuciones de frecuencias agrupadas en intervalos, no es posible identificar directamente al valor central. En este caso seguiremos los pasos siguientes:

- (i) Se calculan las frecuencias acumuladas, N_i .
- (ii) El intervalo mediano es el primer intervalo cuya frecuencia acumulada supere el valor $\frac{N}{2}$. Supongamos que es el $(l_{i-1}, l_i]$.
- (iii) Suponiendo que los valores de la variable se reparten uniformemente dentro de cada intervalo, entonces:

$$M_e = l_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}} \cdot c_i$$

EJEMPLO 2.7 Determinemos el valor mediano para la variable $X = \text{"Gasto total en miles de pesetas de 58 familias"}$

$l_{i-1} - l_i (x1000)$	n_i	N_i
0 - 50	1	1
50 - 100	10	11
100 - 150	9	20
150-200	12	32
200 - 250	12	44
250 - 300	10	54
300 - 350	3	57
350 - 400	1	58

$$M_e = 150 + \frac{29 - 20}{32 - 20} \cdot 50 = 187,5 \Rightarrow M_e = 187\,500 \text{ pesetas}$$

Ventajas de la mediana

- (a) Se puede calcular aún cuando los intervalos inicial y final sean abiertos.
- (b) No se ve afectada por outliers. Es una medida robusta.
- (c) Cuando tengamos distribuciones sesgadas o asimétricas, la mediana es la medida más representativa, ya que ocupa una posición central, entre la media y la moda, en distribuciones normales.

Inconvenientes de la mediana

- (a) No intervienen todos los valores de la variable.
- (b) La mediana es sensible a cualquier cambio en el número de elementos que estemos considerando.
- (c) Desde un punto de vista computacional, ordenar es un proceso lento, es por lo que es preferible el cálculo de la media.

OBSERVACIÓN 2.4 *Siempre será recomendable calcular tanto la media como la mediana: ambas diferirán mucho cuando haya heterogeneidad en los datos (datos con gran dispersión o con presencia de outliers).*

2.1.4. La moda

Definición 2.7 *La moda es el valor de la variable que más se repite.*

OBSERVACIÓN 2.5 *Existen distribuciones que pueden tener más de una moda.*

- (a) Distribuciones no agrupadas en intervalos. La moda es el valor de la variable de mayor frecuencia absoluta.

EJEMPLO 2.8 *Se consideran los datos correspondientes al número de componentes de 75 familias.*

x_i	1	2	3	<u>4</u>	5	6	7	9
n_i	6	11	11	20	15	8	3	1

$M_o = 4 \Rightarrow$ Respuesta: El tamaño de familia más frecuente es de 4 miembros.

- (b) Distribuciones agrupadas en intervalos. Supongamos que todos los valores del intervalo se encuentran distribuidos uniformemente, y que la moda está más cerca de aquel intervalo contiguo cuya frecuencia sea mayor.
 - (i) Si todos los intervalos son de **igual amplitud**, los pasos a seguir son:

- El intervalo modal, $(l_{i-1}, l_i]$, es aquel que tiene la mayor frecuencia absoluta, n_i .
- El valor modal será:

$$M_o = l_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot c_i \quad (c_i = c, \text{ para todo valor de } i)$$

EJEMPLO 2.9 *Se considera el estudio de la variable $X = \text{"Gasto total en miles de pesetas"}$ de 58 familias. Calculemos la moda.*

$l_{i-1} - l_i(x1000)$	n_i	c_i
0 - 50	1	50
50 - 100	10	50
100 - 150	9	50
150 - 200	12	50
200 - 250	12	50
250 - 300	10	50
300 - 350	3	50
350 - 400	1	50

$$M_o = 150 + \frac{12}{9 + 12} \cdot 50 = 178,57 \Rightarrow M_o = 178\,570 \text{ pesetas}$$

$$M'_o = 200 + \frac{10}{12 + 10} \cdot 50 = 222,72 \Rightarrow M'_o = 222\,720 \text{ pesetas}$$

- (ii) Si los intervalos son de **distinta amplitud**, utilizaremos las densidades, $d_i = \frac{n_i}{c_i}$, en vez de las frecuencias absolutas, n_i .

$$M_o = l_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot c_i$$

EJEMPLO 2.10 *Se considera el estudio de la variable $X = \text{"Edad"}$ de un grupo de personas.*

$l_{i-1} - l_i$	n_i	c_i	d_i
16 - 19	9	3	3
19 - 24	40	5	8
24 - 34	50	10	5
34 - 45	22	11	2

$$M_o = 19 + \frac{5}{3 + 5} \cdot 5 = 22,125 \Rightarrow \text{La edad más frecuente es 22 años.}$$

Ventajas de la moda

- (a) No le afectan los outliers.

- (b) Puede usarse para datos cualitativos.
- (c) Es de sencillo cálculo.
- (d) Es fácil de reconocer en los gráficos. En un gráfico de sectores es el sector más grande; en un gráfico de barras, la barra más alta, etc.
- (e) Existen casos en que la media y mediana no son representativas y sí la moda. Esto ocurre cuando los valores de la variable son esencialmente el mismo valor para todos los casos, excepto para unos cuantos.

EJEMPLO 2.11 *En la tabla siguiente se muestra la estructura de sueldo por hora en euros de un restaurante Telepizza:*

<i>Clasificación de empleados</i>	<i>Sueldo</i>	<i>n_i</i>
<i>Empleado regular</i>	<i>5.75</i>	<i>13</i>
<i>Gerente nocturno</i>	<i>10.5</i>	<i>2</i>
<i>Gerente en jefe</i>	<i>18.9</i>	<i>1</i>

La media 7,17 está inflada por los sueldos de los gerentes. Y la mediana 5,75, es igual que la moda, pero lleva a la interpretación incorrecta de que la mitad de los empleados ganan más que esa cantidad, lo cual no es el caso. Decir que la moda es 5,75 significa que muchos empleados se les paga ese sueldo bajo, que es lo más representativo de esta distribución de sueldos.

Inconvenientes de la moda

- (a) No intervienen ni todos los valores como en la media aritmética, ni todas las frecuencias como con la mediana.
- (b) Insensible a los valores de la variable y a cómo se distribuyen a su alrededor los valores de la variable.
- (c) La moda puede ser engañosa cuando se usa sola porque es insensible tanto a outliers como al número de datos que estamos considerando. Esto significa que podemos tener cualquier número de distribuciones con formas totalmente diferentes, y aún todas podrían tener la misma moda.
- (d) No hay moda si todos los valores tienen igual frecuencia.

2.2. Medidas de posición no centrales

Definición 2.8 *Los cuantiles son valores que, una vez ordenada de menor a mayor la distribución, la dividen en partes iguales, es decir, en intervalos que comprenden el mismo número de valores.*

Entre los cuantiles podemos citar:

- (a) Los *cuartiles*. Son tres valores que, una vez ordenada de menor a mayor la distribución, la dividen en cuatro partes iguales. Es decir, en cuatro intervalos dentro de cada uno de los cuales está contenido el 25 % de los valores. Los representaremos por $Q_{r/4}$ con $r = 1, 2, 3$. (Observemos que $Q_{1/2} = M_e$)
- (b) Los *deciles*. Son nueve valores que, una vez ordenada de menor a mayor la distribución, la dividen en diez partes iguales. Dentro de cada una está contenido el 10 % de los valores de la distribución. Los expresaremos como $Q_{r/10}$ con $r = 1, 2, \dots, 9$. ($Q_{5/10} = M_e$)
- (c) Los *percentiles*. Son 99 valores que dividen a la distribución en cien partes iguales, una vez ordenada de menor a mayor. De este modo, entre dos percentiles consecutivos encontramos un 1 % de los datos. Escribiremos $Q_{r/100}$ con $r = 1, 2, \dots, 99$. ($Q_{50/100} = M_e$)

Cálculo de los cuantiles

- (a) En distribuciones de frecuencias sin agrupar.

$$(i) \text{ Si } N_{i-1} < \frac{r}{k} \cdot N < N_i \Rightarrow Q_{r/k} = x_i$$

$$(ii) \text{ Si } N_i = \frac{r}{k} \cdot N \Rightarrow Q_{r/k} = \frac{x_i + x_{i+1}}{2}$$

EJEMPLO 2.12 Consideremos los datos siguientes sobre el tamaño de un conjunto de 75 familias. Calculemos el séptimo decil:

$$N = 75 \Rightarrow \frac{7 \cdot N}{10} = \frac{7 \cdot 75}{10} = 52,5$$

x_i	n_i	N_i
1	6	6
2	11	17
3	11	28
4	20	48
5	15	63
6	8	71
7	3	74
9	1	75

$$Q_{7/10} = 5$$

- (b) En las distribuciones de frecuencias agrupadas en intervalos procederemos de la forma siguiente:
 - (i) Se calculan las frecuencias acumuladas, N_i .

- (ii) El intervalo cuantil es el primer intervalo cuya frecuencia acumulada supere el valor $\frac{r}{k}N$. Supongamos que es el $(l_{i-1}, l_i]$.
- (iii) Suponiendo que los valores de la variable se reparten uniformemente dentro de cada intervalo, entonces:

$$Q_{r/k} = l_{i-1} + \frac{\frac{r}{k} \cdot N - N_{i-1}}{N_i - N_{i-1}} \cdot c_i$$

EJEMPLO 2.13 *Se considera el estudio de la variable $X = \text{"Gasto total en miles de pesetas de 58 familias"}$. Determinemos el primer cuartil, el percentil cuarenta y dos, y el percentil en el que se encuentra un gasto de 125000 pesetas.*

$l_{i-1} - l_i (\times 1000)$	n_i	N_i
0 - 50	1	1
50 - 100	10	11
<u>100-150</u>	9	20
<u>150-200</u>	12	32
200 - 250	12	44
250 - 300	10	54
300 - 350	3	57
350 - 400	1	58

- $\frac{N}{4} = 14,5 \Rightarrow Q_{1/4} = 100 + \frac{14,5 - 11}{20 - 11} \cdot 50 = 119,44 \Rightarrow 119\,440 \text{ ptas.}$
- $\frac{42}{100} \cdot N = 24,36 \Rightarrow Q_{42/100} = 150 + \frac{24,36 - 20}{32 - 20} \cdot 50 = 168,16.$
-

$$125 = 100 + \frac{\frac{r}{100} \cdot 58 - 11}{20 - 11} \cdot 50 \Rightarrow r = 26$$

En ambos casos, para $k = 4$ y $r = 1, 2, 3$, tendremos los cuartiles; para $k = 10$ y $r = 1, 2, \dots, 9$, los deciles, y para $k = 100$ y $r = 1, 2, \dots, 99$, los percentiles.

2.3. Momentos no centrados y centrados

Definición 2.9 *Los momentos de una distribución son valores que la caracterizan.*

OBSERVACIÓN 2.6 *Dos distribuciones son iguales si tienen todos sus momentos iguales.*

- (a) Se definen los *momentos no centrados*, respecto al origen o momentos ordinarios de orden r de la distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$ de la forma siguiente:

$$a_r = \sum_{i=1}^k (x_i - 0)^r \cdot \frac{n_i}{N}$$

$$r = 0, 1, 2, \dots$$

Casos particulares:

$$a_0 = \sum_{i=1}^k x_i^0 \cdot \frac{n_i}{N} = 1; \quad a_1 = \sum_{i=1}^k x_i^1 \cdot \frac{n_i}{N} = \bar{x}; \quad a_2 = \sum_{i=1}^k x_i^2 \cdot \frac{n_i}{N}$$

- (b) Los *momentos centrados* de orden r o momentos respecto de la media de orden r de una distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$ se definen como:

$$m_r = \sum_{i=1}^k (x_i - \bar{x})^r \cdot \frac{n_i}{N}$$

$$r = 0, 1, 2, \dots$$

Casos particulares:

$$m_0 = \sum_{i=1}^k (x_i - \bar{x})^0 \cdot \frac{n_i}{N} = 1; \quad m_1 = \sum_{i=1}^k (x_i - \bar{x})^1 \cdot \frac{n_i}{N} = 0$$

$$m_2 = \sum_{i=1}^k (x_i - \bar{x})^2 \cdot \frac{n_i}{N}; \quad m_3 = \sum_{i=1}^k (x_i - \bar{x})^3 \cdot \frac{n_i}{N}$$

OBSERVACIÓN 2.7 *Todos los momentos centrados se pueden poner en función de los momentos no centrados.*

Como casos particulares de la anterior propiedad se pueden comprobar fácilmente las siguientes igualdades:

$$m_2 = a_2 - a_1^2$$

$$m_3 = a_3 - 3a_2a_1 + 2a_1^3$$

$$m_4 = a_4 - 4a_3a_1 + 6a_2a_1^2 - 3a_1^4$$

Capítulo 3

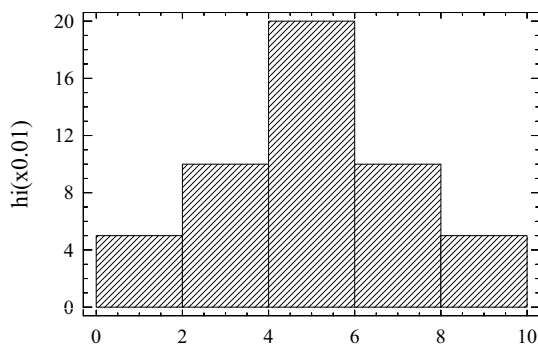
Resumen de datos: Medidas de dispersión

Consideremos dos distribuciones de frecuencias que estudian una misma característica sobre la población, para las cuales se obtiene la misma medida de posición. Veamos como el comportamiento puede ser bien distinto.

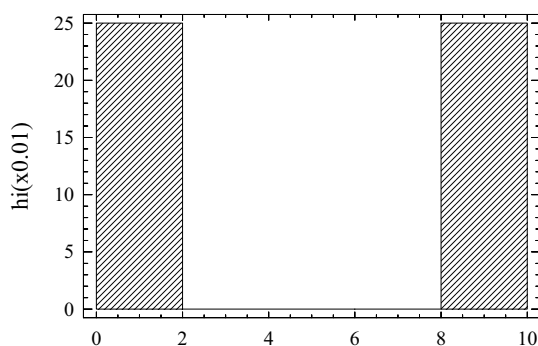
EJEMPLO 3.1 *Se consideran las distribuciones correspondientes a las "calificaciones en cierta asignatura", de 20 alumnos que pertenecen a los grupos A y B, respectivamente. Observemos el distinto comportamiento de estas distribuciones que tienen media aritmética común igual a 5 puntos.*

$l_{i-1} - l_i$	n_i	x_i	$x_i n_i$
0 - 2	2	1	2
2 - 4	4	3	12
4 - 6	8	5	40
6 - 8	4	7	28
8 - 10	2	9	18
	20		100

$l_{i-1} - l_i$	n_i	x_i	$x_i n_i$
0 - 2	10	1	10
8 - 10	10	9	90
	20		100



Grupo A



Grupo B

Definición 3.1 Se llama *dispersión o variabilidad*, a la menor o mayor separación de los valores respecto a otro que se pretende sea la síntesis.

Surgen diferentes medidas de dispersión. Pueden definirse teniendo en cuenta:

- (a) La diferencia entre determinados valores de la variable.
- (b) Promedios de las diferencias entre cada valor de la variable y una medida de posición ($\bar{x}$, M_e , por ejemplo).
- (c) La idea de que no dependa de las unidades de medida de los valores.

3.1. Medidas de dispersión absoluta

Son medidas de dispersión que vienen expresadas en las mismas unidades de medida que los datos.

3.1.1. Recorridos. Desviaciones medias

Definición 3.2 *El recorrido es la diferencia entre el mayor y el menor valor de la variable:*

$$R = \max_i x_i - \min_i x_i$$

Ventajas del recorrido

- (a) Es una medida de fácil cálculo.
- (b) Es útil en situaciones en las que se requiera medir la dispersión con mucha frecuencia y sobre pocos valores.

Inconvenientes del recorrido

- (a) Total dependencia de los valores extremos de la serie de datos. Un outlier hará que su valor sea poco representativo.
- (b) No puede ser calculado si el valor máximo o el mínimo no están determinados.
- (c) No tiene en cuenta los valores intermedios de la variable, así que puede no ser muy preciso.
- (d) No nos dice nada sobre la forma de la distribución entre las puntuaciones extremas. Podemos tener distribuciones con el mismo recorrido y sus formas ser radicalmente diferentes.

Definición 3.3 *El recorrido intercuartílico nos indica el intervalo donde están comprendidos el 50 % central de los valores, y se calcula:*

$$R_I = Q_{3/4} - Q_{1/4}$$

OBSERVACIÓN 3.1 *Presenta como ventaja respecto al recorrido, la eliminación del posible efecto que pudieran tener algunos valores extremos.*

EJEMPLO 3.2 *Si recordamos el ejemplo 3.1, tenemos:*

$$R(A) = \max_i x_i - \min_i x_i = 9 - 1 = 8 ; R(B) = \max_i x_i - \min_i x_i = 9 - 1 = 8$$

$$R_I(A) = Q_{3/4} - Q_{1/4} = 6,5 - 3,5 = 3 ; R_I(B) = Q_{3/4} - Q_{1/4} = 9 - 1 = 8$$

Pero necesitamos medidas de dispersión que involucren a las de posición. La primera solución sería considerar las desviaciones de cada valor con respecto a una medida de posición, p , y promediar posteriormente, estas desviaciones. Es decir:

$$D_p = \sum_{i=1}^k (x_i - p) \cdot \frac{n_i}{N}$$

Esta definición presenta el inconveniente de que si tenemos una distribución muy dispersa a ambos lados de p habrá desviaciones de distinto signo que al sumarse se compensarán, lo que puede hacer que una desviación grande se transforme en una pequeña. Una solución es considerar el valor absoluto o elevar al cuadrado tales desviaciones, y así medir la proximidad o lejanía de los valores, x_i , a la medida de posición empleada.

Definición 3.4 *La desviación media respecto a la media aritmética es la media aritmética de los valores absolutos de las desviaciones de los datos respecto de la media.*

$$D_{\bar{x}} = \sum_{i=1}^k |x_i - \bar{x}| \cdot \frac{n_i}{N}$$

Definición 3.5 *La desviación media respecto a la mediana es la media aritmética de los valores absolutos de las desviaciones de los datos respecto de la mediana.*

$$D_{M_e} = \sum_{i=1}^k |x_i - M_e| \cdot \frac{n_i}{N}$$

EJEMPLO 3.3 *Calculemos los valores de $D_{\bar{x}}$ y D_{M_e} para la distribución siguiente:*

$l_{i-1} - l_i$	n_i	N_i	x_i	$x_i n_i$	$ x_i - \bar{x} \cdot n_i$	$ x_i - M_e \cdot n_i$
50 – 125	11	11	87,5	962,5	783,75	337,48
125 – 200	4	15	162,5	650	15	177,28
200 – 275	2	17	237,5	475	157,5	238,64
275 – 350	1	18	312,5	312,5	153,75	194,32
350 – 425	2	20	387,5	775	457,5	538,64
	20			3175	1 567,5	1 486,36

$$\bar{x} = 158,75 \text{ unidades}$$

$$M_e = 50 + \frac{10 - 0}{11 - 0} \cdot 75 = 118,18 \text{ unidades}$$

$$D_{\bar{x}} = \sum_{i=1}^k |x_i - \bar{x}| \cdot \frac{n_i}{N} = \frac{1\,567,5}{20} = 78,375 \text{ unidades}$$

$$D_{M_e} = \sum_{i=1}^k |x_i - M_e| \cdot \frac{n_i}{N} = \frac{1\,486,36}{20} = 74,318 \text{ unidades}$$

3.1.2. Varianza y desviación típica

Son las dos medidas de dispersión absoluta más importantes.

Definición 3.6 *La varianza es la media aritmética de los cuadrados de las desviaciones de los datos respecto a la media. Es decir,*

$$s^2 = m_2 = \sum_{i=1}^k (x_i - \bar{x})^2 \cdot \frac{n_i}{N}$$

Propiedades de la varianza

- (a) La varianza es siempre mayor o igual que cero, por ser suma de cuadrados, y se anula solamente cuando todos los valores de la variable son iguales entre sí.
- (b) La varianza es la medida cuadrática de dispersión óptima ya que, para cualquier valor p se verifica que:

$$s^2 = \sum_{i=1}^k (x_i - \bar{x})^2 \cdot \frac{n_i}{N} \leq \sum_{i=1}^k (x_i - p)^2 \cdot \frac{n_i}{N}$$

(Teorema de König)

- (c) $s^2 = m_2 = a_2 - a_1^2 = \sum_{i=1}^k x_i^2 \cdot \frac{n_i}{N} - \bar{x}^2$
- (d) Viene expresada en las unidades de la variable elevadas al cuadrado.
- (e) Ya que se calcula a partir de la media, igual que ésta, se deja influir por outliers.

Definición 3.7 *Se define la desviación típica como la raíz cuadrada positiva de la varianza. Es decir:*

$$s = +\sqrt{s^2} = +\sqrt{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot \frac{n_i}{N}}$$

Propiedades de la desviación típica

- (a) Es siempre mayor o igual que cero.
- (b) Es una medida de dispersión óptima.
- (c) Valores pequeños de la desviación típica indican poca dispersión de las observaciones con respecto a la media.
- (d) Ya que se calcula a partir de la media, igual que ésta, se deja influir por outliers.

$$(e) \quad s = +\sqrt{a_2 - a_1^2} = +\sqrt{\sum_{i=1}^k x_i^2 \cdot \frac{n_i}{N} - \bar{x}^2}$$

(f) Viene medida en las mismas unidades de la variable.

(g) El intervalo $(\bar{x} - 2s, \bar{x} + 2s)$ contiene al menos el 75 % de los valores de la distribución.

EJEMPLO 3.4 *Calculemos la desviación típica para los datos del ejemplo 3.3:*

$l_{i-1} - l_i$	n_i	x_i	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{x})^2 \cdot n_i$
50 - 125	11	87,5	962,5	84 218,75	55 842,1875
125 - 200	4	162,5	650	105 625	56,25
200 - 275	2	237,5	475	112 812,5	12 403,125
275 - 350	1	312,5	312,5	97 656,25	23 639,0625
350 - 425	2	387,5	775	300 312,5	104 653,125
	20		3 175	700 625	196 593,75

$$\bar{x} = 158,75 \text{ unidades}$$

$$s^2 = \sum_{i=1}^k (x_i - \bar{x})^2 \cdot \frac{n_i}{N} = \frac{196 593,75}{20} = 9 829,6875 \text{ unidades}^2, \text{ o bien}$$

$$s^2 = a_2 - a_1^2 = \sum_{i=1}^k x_i^2 \cdot \frac{n_i}{N} - \bar{x}^2 = \frac{700 625}{20} - (158,75)^2 = 9 829,6875 \text{ unidades}^2$$

$$s = +\sqrt{s^2} = +\sqrt{9 829,6875} = 99,1448 \text{ unidades}$$

EJEMPLO 3.5 *En la tabla que aparece a continuación, se recogen las medidas que resumen los datos de porcentajes de rentabilidad de estos dos tipos de inversión durante el período considerado.*

	Acciones	Bonos del Tesoro
Media	8.2 %	5.8 %
Desviación Típica	20.6 %	1.4 %

En el contexto de este ejemplo, puede pensarse en la desviación típica como una medida de la incertidumbre o riesgo de la rentabilidad de una inversión. Es decir, la rentabilidad de las acciones fue mayor, pero su riesgo, que viene medido por la desviación típica de la rentabilidad, fue también mayor.

OBSERVACIÓN 3.2 *A la hora de escoger una medida que describa la dispersión de un conjunto de datos, la desviación media respecto de la media aritmética tiene dos ventajas respecto a la desviación típica.*

- (a) *En primer lugar, es más fácil de entender "el promedio de las desviaciones absolutas respecto de la media" que "la raíz cuadrada del promedio del cuadrado de las desviaciones respecto de la media".*
- (b) *Dado que en el cálculo de la varianza y la desviación típica se elevan al cuadrado las desviaciones individuales, estas dos medidas se verán más afectadas por outliers que la desviación media respecto de la media.*

A pesar de estas ventajas, la desviación media respecto de la media se utiliza poco debido a la dificultad algebraica que supone trabajar con el valor absoluto.

Efecto sobre la varianza y la desviación típica de una transformación lineal

A continuación estudiemos cómo se ven afectadas la varianza y la desviación típica ante una transformación lineal que puedan sufrir los datos.

- (a) Dada una distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$, vamos a considerar una nueva distribución $\{(x'_i; n_i)\}_{i=1,2,\dots,k}$, donde $x'_i = x_i + b$ para todo valor de i . Se verifica que:

$$\begin{aligned} s'^2 &= \sum_{i=1}^k (x'_i - \bar{x}')^2 \cdot \frac{n_i}{N} = \\ &= \sum_{i=1}^k [x_i + b - (\bar{x} + b)]^2 \cdot \frac{n_i}{N} = \sum_{i=1}^k [x_i - \bar{x}]^2 \cdot \frac{n_i}{N} = s^2 \end{aligned}$$

Por tanto, si a todos los valores de una variable le sumamos una constante b , la varianza (y la desviación típica) no varían. Es decir, *a la varianza (y a la desviación típica) no le afectan los cambios de origen.*

- (b) Dada una distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$, vamos a considerar una nueva distribución $\{(x'_i; n_i)\}_{i=1,2,\dots,k}$, donde $x'_i = ax_i$ para todo valor de i . Se verifica que:

$$\begin{aligned} s'^2 &= \sum_{i=1}^k (x'_i - \bar{x}')^2 \cdot \frac{n_i}{N} = \\ &= \sum_{i=1}^k [ax_i - (a\bar{x})]^2 \cdot \frac{n_i}{N} = a^2 \cdot \sum_{i=1}^k [x_i - \bar{x}]^2 \cdot \frac{n_i}{N} = a^2 \cdot s^2 \end{aligned}$$

Por tanto, si todos los valores de una variable los multiplicamos por una constante a , la varianza queda también multiplicada por el cuadrado de la constante (y la desviación típica por la propia constante).

Consecuencia

Dada una distribución de frecuencias $\{(x_i; n_i)\}_{i=1,2,\dots,k}$, si consideramos una nueva distribución $\{(x'_i; n_i)\}_{i=1,2,\dots,k}$, que sea una transformación lineal de la primera, es decir, $x'_i = ax_i + b$ para todo valor de i , entonces se verifica que:

$$s'^2 = a^2 s^2 (s' = |a| s)$$

EJEMPLO 3.6 Si tenemos

$$\left. \begin{array}{l} \bar{x} = 10; s^2 = 2 \\ x'_i = 3x_i + 2 \end{array} \right\} \Rightarrow s'^2 = 3^2 s^2 = 9 \times 2 = 18$$

3.1.3. Normalización o tipificación

Desde el punto de vista estadístico, la transformación lineal más importante es la conocida como tipificación o normalización.

Definición 3.8 Dada una variable estadística X , con media $\bar{x}$ y desviación típica s_X , entonces la tipificación consiste en la transformación:

$$z = \frac{x - \bar{x}}{s_X}$$

OBSERVACIÓN 3.3

- (a) Teniendo en cuenta como afectan a la media y a la varianza las transformaciones lineales, se tiene que $\bar{z} = 0$ y $s_z^2 = 1$.
- (b) La variable tipificada expresa el número de desviaciones típicas que cada observación dista de la media. Así podremos comparar la posición relativa de datos de diferentes distribuciones.

EJEMPLO 3.7 Queremos comparar las notas de Estadística Administrativa de dos alumnos, uno que pertenece al grupo A y otro al B. En el grupo A, donde la calificación media fue de 6.2 puntos con una desviación típica de 2.2 puntos, el alumno seleccionado obtuvo 6.8 puntos. El alumno del grupo B obtuvo una calificación de 6, siendo la calificación media de ese grupo 5.2 con una desviación típica de 1.4.

$$z_A = \frac{x_A - \bar{x}_A}{s_A} = \frac{6,8 - 6,2}{2,2} = 0,27$$

$$z_B = \frac{x_B - \bar{x}_B}{s_B} = \frac{6 - 5,2}{1,4} = 0,57$$

Respuesta: El resultado de 6 puntos de la clase B es comparativamente mejor que el de 6.8 en la A, aunque éste sea más alto en términos absolutos.

3.2. Medidas de dispersión relativa

Supongamos que tenemos dos distribuciones de frecuencias cuyas medidas de posición son p_1 y p_2 y queremos saber cuál de las dos es más representativa. Como tales medidas pueden venir expresadas en distintas unidades, no podremos comparar la representatividad de ambas utilizando las medidas de dispersión absoluta.

Es preciso construir medidas de dispersión *adimensionales*, es decir, medidas que resulten independientes de la unidad con que se miden los valores de cada variable. Son las medidas de dispersión relativas.

3.2.1. Recorridos

Definición 3.9 *El recorrido relativo viene dado por la expresión:*

$$R_r = \frac{R}{\bar{x}} = \frac{\max_i x_i - \min_i x_i}{\bar{x}}$$

Nos proporciona el número de veces que el recorrido contiene a la media aritmética.

Definición 3.10 *El recorrido semiintercuartílico se define como:*

$$R_{SI} = \frac{\frac{Q_{3/4} - Q_{1/4}}{2}}{\frac{Q_{1/4} + Q_{3/4}}{2}} = \frac{Q_{3/4} - Q_{1/4}}{Q_{1/4} + Q_{3/4}}$$

Compara la distancia media entre los cuartiles primero y tercero con el punto medio de dicho intervalo.

3.2.2. El coeficiente de variación de Pearson

Definición 3.11 *Se define por la expresión:*

$$V = \frac{s}{|\bar{x}|}$$

Propiedades del coeficiente de variación de Pearson

- (a) Es una medida adimensional y suele expresarse multiplicada por cien, es decir en forma de porcentaje.
- (b) Representa el número de veces que la desviación típica contiene a $|\bar{x}|$. Cuanto mayor es V menos representativa es $\bar{x}$.
- (c) La máxima representatividad de $\bar{x}$ se tiene cuando $V = 0$. Dudaremos de la representatividad de $\bar{x}$ si $V > 0,5$.

(d) Si $\bar{x} = 0$, V no es calculable.

EJEMPLO 3.8 Se consideran los beneficios, expresados en millones de u.m., de dos grupos de empresas, A y B. Estudiemos qué grupo tiene un beneficio medio más representativo.

Grupo de empresas A			
Beneficios en millones de pesetas			
x_i	n_i	$x_i n_i$	$x_i^2 n_i$
1	4	4	4
1.1	6	6.6	7.26
1.2	6	7.2	8.64
1.3	2	2.6	3.38
1.4	2	2.8	3.92
	20	23.2	27.2

Grupo de empresas B			
Beneficios en millones de dólares			
y_j	n_j	$y_j n_j$	$y_j^2 n_j$
1	2	2	2
1.1	2	2.2	2.42
1.2	4	4.8	5.76
1.3	4	5.2	6.76
1.4	4	5.6	7.84
1.5	2	3.0	4.5
1.6	2	3.2	5.12
	20	26.0	34.4

$$\bar{x} = \frac{23,2}{20} = 1,16 \text{ millones de pesetas}; \bar{y} = \frac{26}{20} = 1,3 \text{ millones de dólares}$$

$$s_X^2 = \frac{27,2}{20} - (1,16)^2 = 0,0144; s_Y^2 = \frac{34,4}{20} - (1,3)^2 = 0,03$$

$$s_X = 0,12 \text{ millones de pesetas}; s_Y = 0,1732 \text{ millones de dólares}$$

$$V_X = \frac{s_X}{|\bar{x}|} = \frac{0,12}{1,16} = 0,1034; V_Y = \frac{s_Y}{|\bar{y}|} = \frac{0,1732}{1,3} = 0,1332$$

Respuesta: El beneficio medio de las empresas del grupo A es más representativo.

Capítulo 4

Resumen de datos: Medidas de forma

En nuestra búsqueda de medidas o parámetros que nos suministren información sobre el comportamiento global de una población hemos visto las medidas de posición y de dispersión. Ahora damos un paso más al intentar precisar la forma de la distribución. Las medidas de forma se dirigen a elaborar valores que midan el aspecto de la representación gráfica de la distribución sin necesidad de llevarla a cabo.

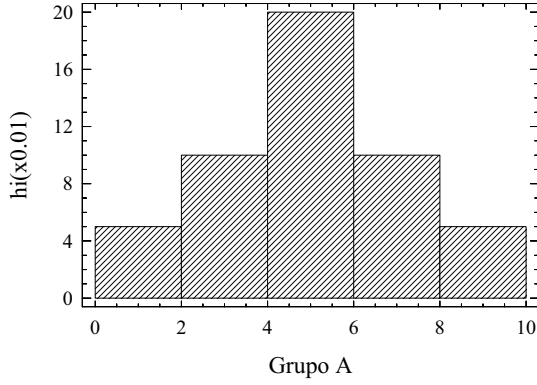
4.1. Medidas de asimetría

Definición 4.1 *Si por $\bar{x}$ trazamos una perpendicular al eje horizontal y la tomamos como eje de simetría, diremos que la distribución es simétrica si existe el mismo número de valores a ambos lados del eje y equidistantes de él hay pares de valores con la misma frecuencia.*

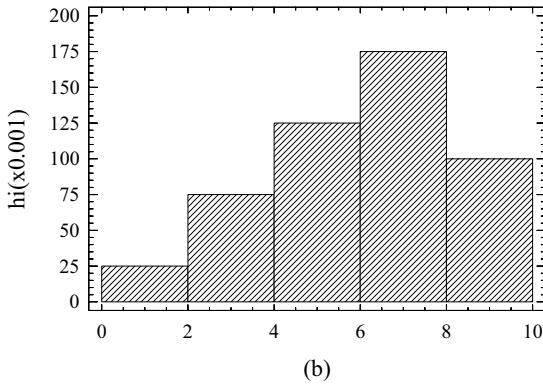
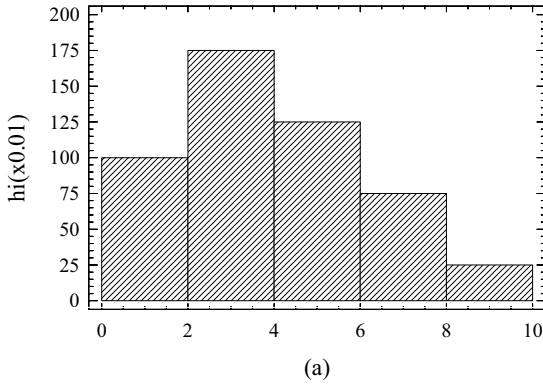
EJEMPLO 4.1 *Teniendo en cuenta el histograma correspondiente al Grupo A del Ejemplo 3.1, observamos que se corresponde con una distribución simétrica respecto de $\bar{x} = 5$.*

Si la distribución no es simétrica, puede ocurrir:

- (a) Que los valores bajos de la variable sean los más frecuentes. Gráficamente observaríamos una cola alargada hacia la derecha y la distribución se dirá que es *asimétrica a derechas* o que tiene asimetría positiva.
- (b) Que los valores altos sean los más frecuentes. Gráficamente se observaría una cola alargada hacia la izquierda y la distribución se dirá que es *asimétrica a izquierdas* o que tiene asimetría negativa.



EJEMPLO 4.2 Mostramos los histogramas de dos distribuciones asimétricas (a) y (b), a derechas e izquierdas, respectivamente.



4.1.1. Coeficiente de asimetría de Fisher

En las distribuciones simétricas, las desviaciones de los valores a la izquierda de la media son igualmente frecuentes que las de los valores a la derecha y, por tanto, todos los momentos centrales de orden impar son nulos.

No podemos considerar como medida de asimetría m_1 puesto que vale cero en cualquier caso. No debemos considerar potencias pares porque nos interesa tener en cuenta el signo de las desviaciones a la media, por tanto, recurriremos a m_3 . Este momento acentúa las desviaciones a la media de los valores altos y bajos de la variable cuando no hay simetría, representando así un índice del sesgo de la distribución.

Definición 4.2 *Se define el coeficiente de asimetría de Fisher como la expresión:*

$$g_1 = \frac{m_3}{s^3} = \frac{\sum_{i=1}^k (x_i - \bar{x})^3}{Ns^3} = \frac{a_3 - 3a_2a_1 + 2a_1^3}{(\sqrt{a_2 - a_1^2})^3}$$

El coeficiente de asimetría de Fisher es una medida adimensional.

- (a) Si la distribución es simétrica $\Rightarrow m_3 = 0 \Rightarrow g_1 = 0$. El recíproco no es cierto, en general, como podemos observar en el ejemplo siguiente tomado de Casas Sánchez y Santos Peña (1996):

EJEMPLO 4.3 *Se considera la distribución siguiente:*

x_i	n_i	$x_i n_i$	$(x_i - \bar{x})^3 \cdot n_i$
0	2	0	-128/729
5/9	3	5/3	3/729
1	1	1	125/729
	6	8/3	0

que, claramente, no es simétrica y para la que $m_3 = 0$ ($\Rightarrow g_1 = 0$).

- (b) Si $g_1 > 0$ debe ser que la distribución está desplazada a la derecha de $\bar{x} \Rightarrow$ Asimétrica a derechas.
- (c) Si $g_1 < 0$ debe ser que la distribución está desplazada a la izquierda de $\bar{x} \Rightarrow$ Asimétrica a izquierdas.

EJEMPLO 4.4 *Calculemos g_1 sobre las distribuciones no simétricas del Ejemplo 4.2,*

(a)

$l_{i-1} - l_i$	n_i	x_i	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{x})^3 \cdot n_i$
0 - 2	4	1	4	4	-108
2 - 4	7	3	21	63	-7
4 - 6	5	5	25	125	5
6 - 8	3	7	21	147	81
8 - 10	1	9	9	81	125
	20		80	420	96

$$\bar{x} = 4 \text{ unidades}; \quad s = \sqrt{\frac{420}{20} - 4^2} = \sqrt{5} = 2,236 \text{ unidades}$$

$$m_3 = \sum_{i=1}^k (x_i - \bar{x})^3 \cdot \frac{n_i}{N} = \frac{96}{20} = 4,8 \text{ unidades}^3$$

$$g_1 = \frac{m_3}{s^3} = \frac{4,8}{2,236^3} = 0,4293 \Rightarrow \text{Asimétrica a derechas}$$

(b)

$l_{i-1} - l_i$	n_i	x_i	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{x})^3 \cdot n_i$
0 - 2	1	1	1	1	-125
2 - 4	3	3	9	27	-81
4 - 6	5	5	25	125	-5
6 - 8	7	7	49	343	7
8 - 10	4	9	36	324	108
	20		120	820	-96

$$\bar{x} = 6 \text{ unidades}; \quad s = \sqrt{\frac{820}{20} - 6^2} = \sqrt{5} = 2,236 \text{ unidades}$$

$$m_3 = \sum_{i=1}^k (x_i - \bar{x})^3 \cdot \frac{n_i}{N} = \frac{-96}{20} = -4,8 \text{ unidades}^3$$

$$g_1 = \frac{m_3}{s^3} = \frac{-4,8}{2,236^3} = -0,4293 \Rightarrow \text{Asimétrica a izquierdas}$$

4.1.2. Coeficiente de asimetría de Pearson

Para distribuciones unimodales y campaniformes se tiene que, en el caso de asimetría a derechas $M_o < M_e < \bar{x}$, y en el caso de asimetría a izquierdas, $\bar{x} < M_e < M_o$. Para distribuciones unimodales, campaniformes y simétricas se verifica que $\bar{x} = M_e = M_o$. Las desigualdades anteriores nos sugieren considerar como medida de asimetría la diferencia entre la media y la moda.

Definición 4.3 Se define el coeficiente de asimetría de Pearson como:

$$A_P = \frac{\bar{x} - M_o}{s}$$

OBSERVACIÓN 4.1 El coeficiente de asimetría de Pearson es una medida adimensional, y debe ser utilizado sólo en distribuciones unimodales y campaniformes. En este caso si la distribución es además simétrica, se tiene que $A_P = 0$. De nuevo, no es cierta la afirmación recíproca.

Se verifica que:

- (a) Si $A_P > 0 \Rightarrow$ la distribución es asimétrica a derechas.
- (b) Si $A_P < 0 \Rightarrow$ la distribución es asimétrica a izquierdas.

OBSERVACIÓN 4.2 Karl Pearson demostró que $\bar{x} - M_o \approx 3(\bar{x} - M_e)$, obteniéndose un llamado segundo coeficiente de asimetría de Pearson dado por la expresión:

$$A'_P = \frac{3(\bar{x} - M_e)}{s}$$

EJEMPLO 4.5 Calculemos A_P sobre las distribuciones consideradas en el ejemplo 4.2.

(a)

$l_{i-1} - l_i$	n_i
0 - 2	4
2 - 4	7
4 - 6	5
6 - 8	3
8 - 10	1
	20

$$\bar{x} = 4 \text{ unidades}; s = 2,236 \text{ unidades}; M_o = 2 + \frac{5}{4+5} \cdot 2 = 3,11 \text{ unidades}$$

$$A_P = \frac{\bar{x} - M_o}{s} = \frac{4 - 3,11}{2,236} = 0,398 \Rightarrow \text{Asimétrica a derechas.}$$

(b)

$l_{i-1} - l_i$	n_i
0 - 2	1
2 - 4	3
4 - 6	5
6 - 8	7
8 - 10	4
	20

$$\bar{x} = 6 \text{ unidades}; s = 2,236 \text{ unidades}; M_o = 6 + \frac{4}{5+4} \cdot 2 = 6,88 \text{ unidades}$$

$$A_P = \frac{\bar{x} - M_o}{s} = \frac{6 - 6,88}{2,236} = -0,393 \Rightarrow \text{Asimétrica a izquierdas.}$$

4.2. Medidas de curtosis

El coeficiente de apuntamiento, aplastamiento, deformación, exceso o curtosis fué introducido por Karl Pearson en 1906.

Definición 4.4 *La curtosis trata de estudiar la mayor o menor concentración de frecuencias alrededor de la media y ver así si la distribución es más o menos apuntada.*

Las medidas de curtosis deben aplicarse sólo a distribuciones unimodales y simétricas, o con ligera asimetría.

Para ello es necesario tener en cuenta una distribución tipo, la distribución Normal. La curva Normal que es simétrica respecto a su media, campaniforme y de recorrido infinito se utiliza como patrón de comparación para el estudio del apuntamiento de una distribución.

4.2.1. Introducción al modelo Normal

La curva representativa de una distribución Normal es la correspondiente a la función matemática dada por:

$$f(x) = \frac{1}{\sqrt{2s\pi}} \cdot e^{-\frac{1}{2} \cdot \left[\frac{x - \bar{x}}{s} \right]^2}$$

y se llama función de densidad de la distribución Normal, donde $\bar{x}$ es la media de la distribución y s la desviación típica.

Esta función tiene, entre otras, las siguientes propiedades:

- (a) Está definida y es positiva para todo valor de x .
- (b) Es continua y derivable en todo su campo de definición.
- (c) Es simétrica respecto a la recta $x = \bar{x}$.
- (d) El eje horizontal es asíntota de la curva.
- (e) Posee un máximo relativo en el punto $\left(\bar{x}, \frac{1}{\sqrt{2s\pi}} \right)$.
- (f) Tiene puntos de inflexión para los valores de x dados por $x = \bar{x} \pm s$.

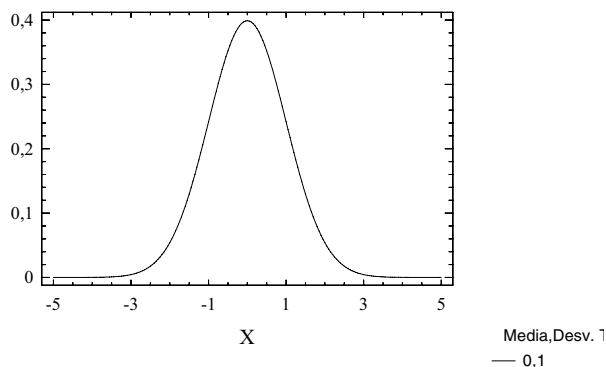


Figura 4.1: Representación gráfica de una distribución Normal

Definición 4.5 Diremos que una distribución es igual de apuntada que la Normal o mesocúrtica ("meso"=mitad), más apuntada que la Normal o leptocúrtica ("lepto"=esbelto), o menos apuntada que la distribución Normal o platicúrtica ("plato"=plano, ancho).

4.2.2. El coeficiente de curtosis de Fisher

El momento central de cuarto orden constituye una medida del apuntamiento de las distribuciones, pues acentúa las desviaciones a la media de los valores situados a la derecha y a la izquierda de ella. Para que la medida sea adimensional se divide por s^4 . Karl Pearson observó que en la distribución Normal se cumple que $\frac{m_4}{s^4} = 3$.

Definición 4.6 El coeficiente de curtosis de Fisher tiene como expresión

$$g_2 = \frac{m_4}{s^4} - 3$$

Se verifica que:

- (a) Si $g_2 > 0 \Rightarrow$ Distribución leptocúrtica.
- (b) Si $g_2 < 0 \Rightarrow$ Distribución platicúrtica.
- (c) Si $g_2 = 0 \Rightarrow$ Distribución mesocúrtica.

EJEMPLO 4.6 Calculemos el coeficiente g_2 para las siguientes distribuciones:

(a)

$l_{i-1} - l_i$	n_i	x_i	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{x})^4 \cdot n_i$
-1 - 1	1	0	0	0	243,8820
1 - 3	22	2	44	88	319,2743
3 - 5	44	4	176	704	0,0002
5 - 7	11	6	66	396	193,5896
7 - 9	4	8	32	256	1 074,2561
9 - 11	1	10	10	100	1 338,1493
	83		328	1 544	3 169,1515

$$\bar{x} = 3,9518 \text{ unidades}; \quad s = \sqrt{\frac{1\,544}{83} - (3,9518)^2} = 1,7279 \text{ unidades}$$

$$m_4 = \sum_{i=1}^k (x_i - \bar{x})^4 \cdot \frac{n_i}{N} = \frac{3\,169,1515}{83} = 38,1825 \text{ unidades}^4.$$

$$g_2 = \frac{m_4}{s^4} - 3 = \frac{38,1825}{1,7279^4} - 3 = 1,2834 \Rightarrow \text{Leptocúrtica}.$$

(b)

$l_{i-1} - l_i$	n_i	x_i	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{x})^4 \cdot n_i$
-1 - 1	5	0	0	0	3 382,6005
1 - 3	16	2	32	64	1 477,6336
3 - 5	22	4	88	352	32,2102
5 - 7	20	6	120	720	13,1220
7 - 9	18	8	144	1 152	1 273,1058
9 - 11	6	10	60	600	3 458,8806
	87		444	2 888	9 637,5527

$$\bar{x} = 5,10 \text{ unidades}; \quad s = \sqrt{\frac{2\,888}{87} - (5,1)^2} = \sqrt{7,1854} = 2,6806 \text{ unidades}$$

$$m_4 = \sum_{i=1}^k (x_i - \bar{x})^4 \cdot \frac{n_i}{N} = \frac{9\,637,5527}{87} = 110,7765 \text{ unidades}^4.$$

$$g_2 = \frac{m_4}{s^4} - 3 = \frac{110,7765}{2,6806^4} - 3 = -0,8545 \Rightarrow \text{Platicúrtica}.$$

OBSERVACIÓN 4.3 El coeficiente de curtosis nos informa de la posible heterogeneidad de los datos. Si es muy bajo indica una distribución mezclada; si es muy alto indica la posible presencia de outliers.

4.3. Análisis Exploratorio de Datos

Esta parte de la Estadística es un conjunto de técnicas, fundamentalmente gráficas, que pretenden dar una visión intuitiva y simple de las principales características de la distribución en estudio. Se analizan exhaustivamente los datos y se detectan posibles anomalías de los mismos. Como ejemplos del Análisis Exploratorio de Datos veremos el diagrama de tallo y hojas y el diagrama de caja y bigotes.

4.3.1. Diagrama de caja y bigotes

El *diagrama de caja y bigotes* (Box&Whisker), introducido por Tukey en el año 1977, se trata de una síntesis gráfica de una distribución en la que intervienen cinco valores: mediana, cuartiles primero y tercero, y los valores máximo y mínimo. Su construcción es como sigue:

- (a) Tomamos una escala que contenga el rango o recorrido de la variable.
- (b) Se dibuja una caja, rectángulo, que vaya desde el primer hasta el tercer cuartil.
- (c) Dibujamos en el interior de la caja una barra vertical en la posición de la mediana.
- (d) Calcular los límites admisibles, superior e inferior, para identificar los outliers

$$LI = Q_{1/4} - 1,5(Q_{3/4} - Q_{1/4}) \quad LS = Q_{3/4} + 1,5(Q_{3/4} - Q_{1/4})$$

- (e) Declarar y marcar como outliers los datos fuera del intervalo (LI,LS).
- (f) Dibujar un bigote o línea que vaya desde cada extremo de la caja hasta el valor más alejado no atípico, es decir, dentro de (LI,LS).

OBSERVACIÓN 4.4

- (a) *Con esta representación se consigue una impresión rápida de ciertas características básicas de un conjunto de datos: posición, dispersión y simetría o asimetría.*
- (b) *La caja del diagrama contiene la mitad central de los datos.*
- (c) *A medida que la mediana esté más centrada en la caja, y cuanto más similares sean las distancias de la caja hasta los valores mínimo y máximo, menos asimétrica es la distribución.*
- (d) *Sirve para identificar los outliers.*

- (e) Suele ser especialmente útil para comparar la distribución de una variable en distintas poblaciones y como un primer acercamiento a razonamientos inferenciales.

EJEMPLO 4.7 Dada la siguiente distribución:

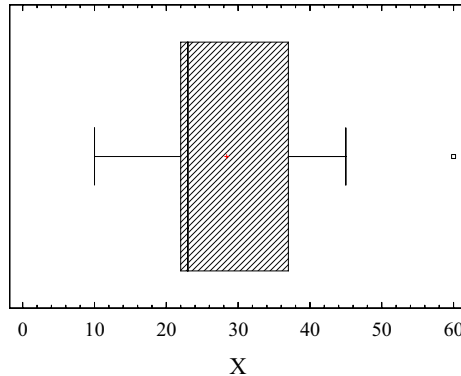
x_i	10	11	22	23	34	40	45	60
n_i	1	3	2	5	4	2	2	1
N_i	1	4	6	11	15	17	19	20

Calculamos $Q_{1/4} = 22$, $Q_{2/4} = 23$ y $Q_{3/4} = 37$. Como en este caso los límites admisibles son:

$$LI = Q_{1/4} - 1,5(Q_{3/4} - Q_{1/4}) = 22 - 1,5(37 - 22) = -0,5$$

$$LS = Q_{3/4} + 1,5(Q_{3/4} - Q_{1/4}) = 37 + 1,5(37 - 22) = 59,5$$

Luego el valor 60 lo declaramos outlier al caer fuera de esos límites. Y por tanto, con el $\max_i x_i = 45$ y el $\min_i x_i = 10$ dentro de los límites admisibles, dibujamos definitivamente el diagrama de caja y bigotes.



4.3.2. Diagrama de tallo y hojas

El *Diagrama de tallo y hojas* o *Histograma digital* (Steam and Leaf plot), introducido también por Tukey y que permite obtener simultáneamente una distribución de frecuencias de la variable y su representación gráfica.

Seguidamente, la construcción del diagrama de tallo y hojas es como sigue:

- Se ordenan en sentido creciente los datos.
- Fijamos los límites admisibles superior e inferior para localizar los outliers.

- (c) Trazamos una línea vertical, que servirá para separar los tallos (escritos a la izquierda) de las hojas (escritas a la derecha).
- (d) Anotamos cada dato en la fila correspondiente al tallo que definen sus primeras cifras. A la derecha de la línea vertical se escribirá su hoja (última cifra).
- (e) A la izquierda del diagrama escribiremos las frecuencias absolutas acumuladas en orden ascendente hasta alcanzar $N/2$, y en orden descendente a continuación. La presencia del valor que ocupa el lugar medio, mediana, en ese tallo queda reflejada por un paréntesis.

EJEMPLO 4.8 Con los datos del ejemplo anterior 4.7

TALLO		HOJAS
4	1	0111
(7)	2	2233333
9	3	4444
5	4	0055
ALTO		60

Ventajas del diagrama de tallo y hojas

- (a) Proporciona las características esenciales de un histograma girando la figura 90° . Nos ayuda a estudiar la forma y simetría de la distribución.
- (b) La condensación de la información respecto a los datos originales es reversible.
- (c) Se suelen identificar los puntos outliers de acuerdo a la regla que hemos expuesto para el diagrama de caja y bigotes.

Inconvenientes del diagrama de tallo y hojas

- (a) Si el número de datos es elevado el método no es operativo.

Capítulo 5

Medidas de desigualdad

5.1. Medidas de desigualdad o concentración

Aún cuando concentración y dispersión tienen significados opuestos, el significado estadístico no coincide con el que corrientemente se da a estos vocablos. La dispersión, estadísticamente hablando, hace referencia a la mayor o menor representatividad de los promedios.

Definición 5.1 *La desigualdad o concentración trata de poner de relieve el mayor o menor grado de igualdad en el reparto de la suma total de los valores de la variable. Las medidas de concentración son, por tanto, indicadores del grado de equidistribución de la variable.*

Las medidas de desigualdad se utilizan, fundamentalmente, para el estudio de variables que se refieren a ingresos, bienes, salarios, deudas, etc. Es decir, tienen una aplicación básicamente económica.

Si consideramos $\sum_{i=1}^k x_i n_i$ el total de la variable a repartir y, a proporciones iguales de la población corresponden proporciones iguales del total de la variable, la distribución de la misma será equitativa. A medida que, a proporciones dadas de la población correspondan proporciones distintas del total, tendremos concentración.

Así diremos que hay

- (a) *Concentración mínima o equidistribución:* si a cada individuo o unidad de la población le corresponde la misma parte del valor total de la variable.
- (b) *Concentración máxima o mínima igualdad en el reparto:* si el total de la variable está concentrada en un sólo individuo o unidad.

5.2. Estudio gráfico: la curva de Lorenz

Dada la variable, X , consideremos su distribución $\{(x_i; n_i)\}_{i=1,2,\dots,k}$. Para construir la curva de Lorenz vamos a organizar nuestros datos ordenados de manera creciente como sigue:

x_i	n_i	N_i	$x_i n_i$	$u_i = \sum_{j=1}^i x_j n_j$	$p_i = \frac{N_i}{N} \cdot 100$	$q_i = \frac{u_i}{u_k} \cdot 100$
x_1	n_1	N_1	$x_1 n_1$	$u_1 = x_1 n_1$	p_1	q_1
x_2	n_2	N_2	$x_2 n_2$	$u_2 = x_1 n_1 + x_2 n_2$	p_2	q_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	N_k	$x_k n_k$	$u_k = x_1 n_1 + \dots + x_k n_k$	p_k	q_k

donde,

- Los productos $x_i n_i$ indican el valor total de la variable para las n_i unidades con valor x_i .
- Las frecuencias acumuladas N_i , indican la cantidad acumulada de unidades.
- Los valores $u_i = \sum_{j=1}^i x_j n_j$, representan el total de la variable acumulado por aquellas unidades con valor menor o igual a x_i .
- Los valores p_i representan la *proporción acumulada de unidades, expresada en porcentajes*.
- Los valores q_i indican la *proporción acumulada del total de la variable, expresada en porcentajes*.

OBSERVACIÓN 5.1 *Las dos últimas columnas nos dan información sobre la concentración.*

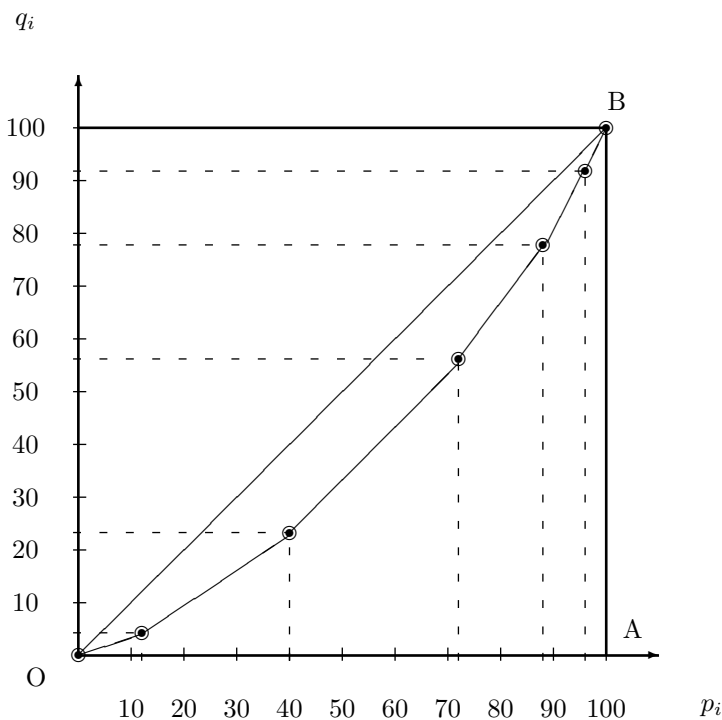
EJEMPLO 5.1 *Analicemos la concentración de los salarios por hora, expresados en miles de unidades monetarias de los trabajadores de una determinada empresa:*

$l_{i-1} - l_i$	x_i	n_i	N_i	$x_i n_i$	$u_i = \sum_{j=1}^i x_j n_j$	$p_i = \frac{N_i}{N} \cdot 100$	$q_i = \frac{u_i}{u_k} \cdot 100$
0,5 – 1,5	1	3	3	3	3	12	4,1
1,5 – 2,5	2	7	10	14	17	40	23,3
2,5 – 3,5	3	8	18	24	41	72	56,2
3,5 – 4,5	4	4	22	16	57	88	78,1
4,5 – 5,5	5	2	24	10	67	96	91,8
5,5 – 6,5	6	1	25	6	73	100	100

OBSERVACIÓN 5.2 Podemos observar como el 12 % de los trabajadores que menos sueldo perciben, acumulan el 4.1 % del total de los salarios. Al 40 % de los trabajadores que menos ganan les corresponde el 23.3 % del total de los salarios, etc... Por tanto, podemos afirmar que no existe equidistribución pues, en ese caso, al 12 % de los trabajadores que menos sueldo perciben les correspondería el 12 % del total de los salarios, al 40 % de los trabajadores el 40 % del total de los salarios y así sucesivamente. Luego, existe cierta concentración.

Esto puede materializarse gráficamente obteniendo la curva definida por la aplicación $p_i \rightarrow q_i$.

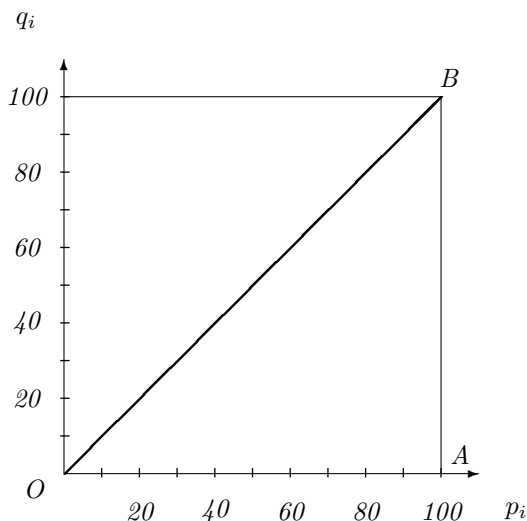
Definición 5.2 La curva de Lorenz es la poligonal que une el punto $(0,0)$ con los puntos (p_i, q_i) para $i = 1, 2, \dots, k$. El nombre le viene del estadístico norteamericano Max Otto Lorenz quien la introdujo en 1905.



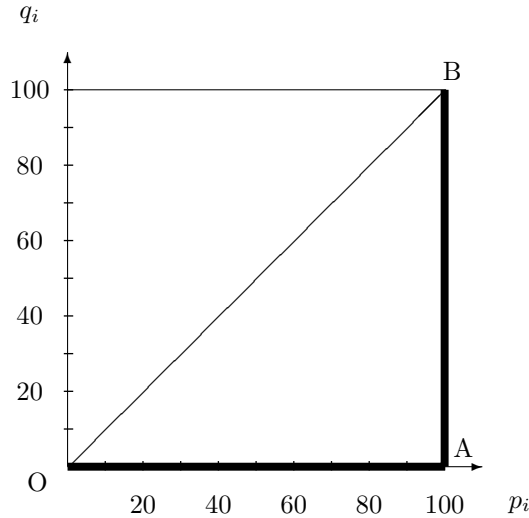
OBSERVACIÓN 5.3

- (a) Como los valores p_i y q_i están comprendidos entre 0 y 100, los puntos (p_i, q_i) estarán en el interior de un cuadrado de lado 100.

- (b) Por la forma en que están definidos los valores p_i y q_i , se tiene que $q_i \leq p_i$ para todo i .
- (c) La curva de Lorenz tiene carácter creciente porque considera porcentajes obtenidos de totales acumulados.
- (d) La curva que nos indicaría la concentración mínima o equidistribución coincidiría con la diagonal $\overline{OB}$ del cuadrado (segmento de equidistribución), ya que en ella $p_i = q_i$ para todo valor de i .



- (e) Si la concentración es máxima, los puntos serán de la forma $(p_i, 0)$, situados sobre el eje de abscisas, salvo el último punto, $(100, 100)$ y la curva de Lorenz coincidiría con los lados del cuadrado situados bajo la diagonal, es decir, los lados $\overline{OA}$ y $\overline{AB}$. Cuanto más se aparte la curva de Lorenz del segmento de equidistribución, tanto mayor será la concentración de la distribución.



5.3. Estudio analítico: el índice de Gini

Basándonos en la observaciones realizadas anteriormente en la construcción de la curva de Lorenz como medida del grado de la desigualdad de una distribución, vamos a estudiar la medida analítica más usual de desigualdad, el índice de Gini.

Observamos que, cuanto más cerca se encuentre la curva de Lorenz del segmento de equidistribución, menor será la concentración de los valores de la variable. Por tanto, la idea consiste en encontrar un coeficiente que intente medir la mayor o menor proximidad de la curva de Lorenz a la diagonal $\overline{OB}$ del cuadrado. Tenemos así una primera definición del índice de Gini.

Definición 5.3 *El índice de Gini es el cociente entre el área de concentración (área comprendida entre la curva de Lorenz y el segmento de equidistribución $\overline{OB}$) y el área máxima de concentración (área de la mitad del cuadrado).*

OBSERVACIÓN 5.4 *Este índice tiene el inconveniente de tener que calcular el área de una figura irregular limitada por una curva que es conocida nada más que en unos cuantos puntos.*

Teniendo en cuenta la observación realizada, se propone otra definición del índice de Gini para una distribución de frecuencias, dada por la expresión:

$$I_G = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i}$$

y que utiliza las distancias horizontales de los puntos (p_i, q_i) al segmento de equidistribución $\overline{OB}$ como medida de la separación de la curva de Lorenz al segmento de equidistribución $\overline{OB}$.

OBSERVACIÓN 5.5

- (a) Se tiene que $0 \leq I_G \leq 1$.
- (b) Si hay equidistribución $\Rightarrow p_i = q_i$ para todo $i \Rightarrow I_G = 0$.
- (c) Si la concentración es máxima $\Rightarrow q_i = 0$ para $i = 1, 2, \dots, k-1 \Rightarrow I_G = 1$.

EJEMPLO 5.2 Calculemos el índice de Gini del ejemplo inicialmente planteado.

$$I_G = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{4,1 + 23,3 + 56,2 + 78,1 + 91,8}{12 + 40 + 72 + 88 + 96} = 1 - \frac{253,5}{308} = 0,1769$$

Entonces hay una baja concentración de los salarios para ese grupo de trabajadores.

Cuestiones, problemas y recursos

Cuestiones

1. Se estudia la variable $X = \text{"número de hijos"}$ a 50 familias, obteniéndose la siguiente tabla:

x_i	0	1	2	3	4	5
n_i	—	15	—	5	4	—
f_i	0,2	—	—	—	—	0,02

Completa los valores que faltan en la tabla y determina el porcentaje de las familias que tienen 3 ó más hijos.

2. Las calificaciones finales de curso obtenidas por los 23 alumnos de una clase fueron:

$Susp - Sob - A - NP - Susp - N - Susp - A - A - A - NP - N$
 $N - A - Susp - Susp - A - A - N - Susp - Susp - Susp - NP$

Construye la distribución de frecuencias correspondiente a los datos y realiza el diagrama de rectángulos correspondiente.

3. Realiza una representación gráfica adecuada de la variable cuya distribución de frecuencias viene dada por la siguiente tabla:

x_i	0	1	2	3	4
n_i	10	8	3	8	10

A la vista de la gráfica, ¿qué tipo de asimetría presenta la distribución? Razona la respuesta.

4. Una fábrica de coches desea estudiar el consumo de un nuevo modelo que quiere lanzar al mercado. Para ello seleccionó 20 coches y midió la distancia en kilómetros que recorrió cada vehículo con diez litros de gasolina. Los resultados fueron:

85.9	88.6	86.2	86.2	86.1	87.3	89.2	88.5	87.4	87.4
88.9	88.9	87.8	87.6	85.8	86.0	86.2	85.9	86.5	86.9

Construye el diagrama de tallo y hojas correspondiente. Calcula la mediana a partir de él e interprétala.

5. Un profesor de cierta asignatura imparte la misma entre 170 alumnos que se reparten entre los grupos A (35 alumnos), B (45 alumnos), C (40 alumnos) y D (50 alumnos) . Las calificaciones medias en cada grupo fueron 7, 6.1, 5.5 y 4.7 puntos, respectivamente. Dicho profesor comentó entre sus compañeros que, entre sus 170 alumnos, la calificación media había sido de

$$\frac{7 + 6,1 + 5,5 + 4,7}{4} = 5,825 \text{ puntos}$$

¿Es correcta tal afirmación? Razona la respuesta proporcionando el valor correcto de la puntuación media si este fuese incorrecto.

6. Dados los valores 2, 4, 5, 8, 1, 3, 3, 6, 150, 8, 6. ¿Cuál de los promedios, media aritmética o mediana, representa mejor a dichos valores? Razona la respuesta.
7. Las ventas efectuadas en el último ejercicio por los vendedores de una empresa han sido, en millones de pesetas, las siguientes: 45, 54, 62, 41, 39, 73, 68, 48, 53, 70, 49, 56. Calcula el tercer cuartil e interpreta el resultado.
8. Se ha estudiado la inversión en publicidad realizada por 150 empresas de un determinado sector económico, obteniéndose que $Q_{20/100} = 80$ millones de pesetas y $Q_{65/100} = 140$ millones de pesetas. Realiza una interpretación adecuada de este resultado.
9. Se han seleccionado dos alumnos que se examinaron de cierta asignatura durante el curso 99/00, encontrándose que ocupan el percentil 40 y 65, respectivamente, en cuanto a la calificación obtenida. Si aprobaron la signatura el 60 % de los alumnos que se presentaron al examen, comenta razonadamente el significado de ambos percentiles y concluye cuál es la situación de estos alumnos respecto a la citada asignatura.
10. ¿Cuál es la utilidad de las medidas de dispersión? Ventajas de las medidas de dispersión relativas frente a las medidas de dispersión absolutas. Ejemplo de una medida de cada tipo.

11. Dadas las siguientes series de datos

(a) 13, 7, 8, 4, 16, 11, 19, 6

(b) 10, 4, 9, 9, 10, 9, 10, 19

Calcula el rango de ambas distribuciones. Realiza algún comentario sobre la medida calculada y la dispersión de ambas series de datos.

12. Las calificaciones obtenidas en matemáticas por dos grupos de alumnos han dado como resultado:

$$\text{Grupo 1} \quad \bar{x} = 5,5 \quad s_X = 0,7$$

$$\text{Grupo 2} \quad \bar{y} = 6,3 \quad s_Y = 0,5$$

¿Qué grupo tiene una media más representativa? Razona la respuesta.

13. Las tasas de desempleo a lo largo de los últimos 15 meses, según informó el Departamento de Trabajo de cierto país, fueron: 4.3, 4.7, 5.4, 5.5, 5.5, 5.6, 5.7, 5.3, 5.7, 6.1, 6.5, 7.2, 7.0, 7.1 y 6.8 %. Realiza el diagrama de caja y bigotes correspondiente. ¿Se trata de una distribución simétrica?

14. Haz la representación gráfica de una distribución unidimensional continua con $A_p = -0,46$ y $g_2 = 4$.

15. El salario medio de 750 trabajadores de una empresa es 250,000 pesetas, con una desviación típica de 11,000 pesetas. A la hora de renovar el convenio colectivo de revisión salarial la dirección de la empresa presenta dos alternativas:

(a) Un incremento proporcional del 20 %.

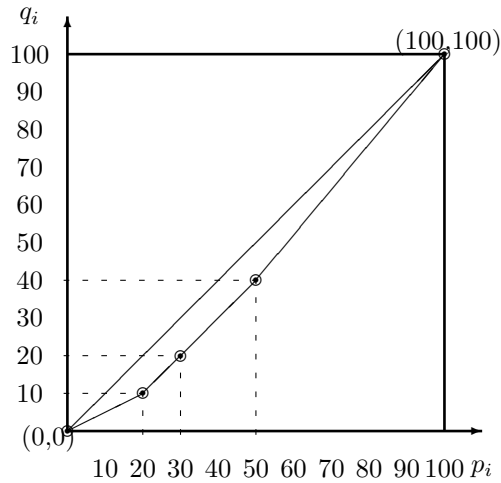
(b) Un aumento fijo de 50,000 pesetas.

¿Con cuál de las dos alternativas habrá mayor disparidad en los salarios?

16. ¿En qué consiste la tipificación? Propiedades que presenta una variable tipificada. Comenta con un ejemplo en qué casos es conveniente tipificar los valores de una variable.

17. Se sabe que el consumo medio de gasolina de una marca de automóviles A es de 7 litros con una desviación típica de 0.7 litros. Para una segunda marca B los valores del consumo medio y la desviación típica fueron 6.5 y 0.8, respectivamente. Determina si un automóvil de la marca B que consumiera 8 litros tendría un consumo relativo mayor a otro de la marca A que consumiera igual. Razona la respuesta.

18. Determina e interpreta el índice de Gini correspondiente al problema siguiente: "Se reparten 10 millones de pesetas entre 5 socios, tres de ellos reciben 2 millones cada uno y el socio que más recibe se lleva 3 millones".
19. Dada la curva de Lorenz siguiente:



determina el índice de Gini correspondiente y comenta el grado de desigualdad en el reparto de los valores de la variable analizada. Relaciona dicho valor con la gráfica de la curva de Lorenz.

Problemas

1. Las calificaciones de 100 estudiantes de una Escuela Superior, *A*, en cierta asignatura fueron las siguientes:

11	46	58	25	38	18	41	35	9	28
35	2	37	68	70	31	44	84	14	82
26	32	51	29	59	92	26	5	52	8
2	12	21	6	32	15	67	37	21	47
33	33	48	47	33	69	49	21	9	15
11	22	29	14	31	36	19	19	21	71
52	32	51	44	57	60	43	65	73	62
3	17	39	22	40	65	30	31	16	80
41	59	60	31	31	10	63	41	74	81
20	36	59	38	40	33	18	60	71	44

- Presenta dichos valores en una tabla estadística adecuada.
- Realiza las siguientes representaciones gráficas: histograma y el polígono de frecuencias ordinarias.
- Calcula la calificación media.
- Calcula la calificación menor conseguida por el 50 % de los alumnos con mayor calificación.
- Calcula la calificación más frecuente.
- Calcula el primer cuartil y el cuarto decil.
- Determina qué calificación tiene como mínimo un alumno para estar dentro del 25 % que más puntuación ha obtenido.
- Los alumnos con al menos una calificación de setenta puntos son considerados para recibir cierta ayuda al estudio, ¿qué porcentaje de alumnos serán considerados?
- ¿Puede considerarse representativa la calificación media?
- Estudia el tipo de asimetría que presenta la distribución.
- Representa el diagrama de caja y bigotes.
- Estudia la curtosis de la distribución.
- En otra Escuela Superior, *B*, se examinó a una serie de alumnos de la misma asignatura obteniéndose una calificación media de 54 puntos con una desviación típica de 30 puntos. Si un alumno obtuvo una calificación de 60 puntos, ¿en qué Escuela Superior destacaría más su calificación?

2. Dada la siguiente distribución de frecuencias, calcula:

x_i	0	1	2	3	4
n_i	1	5	3	1	2

- (a) La media y su representatividad.
- (b) La moda y la mediana.
- (c) El coeficiente de asimetría de Fisher. Interpretación.
- (d) El coeficiente de curtosis. Interpretación.

3. Dada la siguiente distribución de frecuencias, calcula:

x_i	0	10	20	30	40
n_i	2	4	7	5	2

- (a) La media y su representatividad.
- (b) La moda y la mediana.
- (c) El coeficiente de asimetría de Fisher. Interpretación.
- (d) El coeficiente de curtosis. Interpretación.

4. En el siguiente cuadro se refleja el número de días de vida útil de 100 envases de leche fresca PULEVA, en condiciones normales de conservación:

X=Número de días	Número de envases
1	8
2	15
3	33
4	24
5	14
6	6

- (a) Obtener la duración media y su representatividad.
- (b) Obtener la duración que más se da.
- (c) Calcula la mediana y su representatividad a través de la desviación media respecto de la mediana.
- (d) Calcula el coeficiente de asimetría de Fisher y el de curtosis y sus interpretaciones.

5. En una compañía de seguros agrarios, el pago de indemnizaciones durante el año 1998 tuvo la siguiente distribución:

Cuantía de las indemnizaciones (en millones de u.m.)	Número de pagos (en cientos)
0-20	10
20-40	12
40-50	30
50-100	40
100-150	10

- Representa adecuadamente la distribución. Calcula la cuantía de la indemnización pagada más frecuentemente por la compañía.
 - Estudia y comenta el tipo de asimetría de la distribución. Relaciona este resultado con la representación obtenida en el apartado (a).
 - Calcula la cuantía mínima de las indemnizaciones que se encuentran entre el 30 % de las más elevadas.
 - Calcula el porcentaje de los pagos inferiores a 49 millones de u.m.
6. En el distrito I de cierta ciudad se ha constatado que las familias residentes se han distribuido, según el tamaño, de la forma siguiente:

Tamaño familiar	Número de familias
0-2	110
2-4	200
4-6	90
6-8	75
8-10	25

- Calcula el número medio de personas por familia.
 - Obtén el tamaño de familia más frecuente.
 - Si sólo se diera la tarjeta de residente al 50 % de las familias de mayor tamaño, ¿qué tamaño debería tener como mínimo una familia para tener derecho a solicitar dicha tarjeta?
 - Responde a la pregunta anterior, si para tener la tarjeta es necesario estar entre el 25 % de las familias de mayor tamaño.
 - ¿Se pueden hacer previsiones fiables en base al tamaño familiar medio?
 - Calcula e interpreta el coeficiente de asimetría de Pearson.
 - Contruye el histograma y el polígono de frecuencias.
7. Las puntuaciones obtenidas por un grupo de alumnos del colegio La Salle-S. José, en un test de psicomotricidad, han sido las siguientes:

Puntuaciones	Número de alumnos
5-10	3
10-15	6
15-20	13
20-25	7
25-30	2

- Calcula la puntuación media y su representatividad a través del coeficiente de variación.
 - Representa el diagrama de caja y bigotes.
 - ¿En qué percentil está situado un alumno que ha obtenido una puntuación de 12? Interpreta el resultado.
 - Coeficiente de asimetría de Pearson. Interpretación.
8. La distribución de las acciones de una determinada sociedad viene agrupada en intervalos según la siguiente tabla:

Número de acciones	Número de accionistas
0-20	10
20-28	32
28-32	50
32-48	8

- Calcula el número de acciones por accionista más frecuente. Propón y calcula una medida que sirva para estudiar su representatividad (ayúdate, para ello, de otras medidas de dispersión estudiadas).
 - Estudia el tipo de asimetría de la distribución.
 - ¿Qué porcentaje de accionistas poseen más de 30 acciones? ¿Y entre 30 y 40?
9. Un empresario monta una pequeña empresa de ventas de vino de Jerez a través de Internet en la que trabajan 10 asalariados cuyos salarios mensuales medios en euros son los que aparecen en la tabla siguiente:

Salarios	Número de trabajadores
1250	2
1400	3
2500	1
4000	2
6500	2

- Construir la curva de Lorenz. Interpretación.

- (b) Construir el índice de Gini. Interpretación.
 - (c) Representar el diagrama de caja y bigotes.
 - (d) Coeficiente de asimetría de Pearson. Interpretación.
10. La siguiente distribución nos muestra los salarios mensuales que perciben en cientos de euros los trabajadores de una empresa:

Salarios	Número de trabajadores
15-25	20
25-35	10
35-45	5
45-55	5

- (a) Construir la curva de Lorenz. Interpretación.
 - (b) Construir el índice de Gini. Interpretación.
 - (c) Salario medio y su representatividad a través del coeficiente de variación.
 - (d) ¿Qué porcentaje de trabajadores ganan menos de 4000 euros?
 - (e) Coeficiente de asimetría de Pearson. Interpretación.
11. En cierto sector productor de bienes de consumo, la inversión en publicidad se distribuye entre sus empresas según la forma siguiente:

Inversión (millones de u.m.)	Número de empresas
65-75	10
75-85	15
85-95	40
95-105	25
105-125	10

- (a) ¿Puede considerarse representativa la inversión media en publicidad?
 - (b) ¿Cuál es la inversión mínima en publicidad del conjunto formado por el 15 % de las empresas que más invierten en publicidad?
 - (c) Estudia la asimetría de la distribución.
 - (d) Realiza un estudio, tanto analítico como gráfico, de la concentración de las inversiones en publicidad.
 - (e) Para el próximo año se pronostica un incremento de un 10 % en las cantidades invertidas en publicidad, ¿cómo afecta esto a la inversión media? ¿Cómo afecta a su representatividad? ¿Qué efecto tendría sobre la concentración?
12. El volumen anual de las exportaciones totales de una provincia española a 31 de diciembre de cierto año es el reflejado en la tabla siguiente:

Cientes	Volumen de las exportaciones
Estados Unidos	1
Francia	1.1
Alemania	2
Reino Unido	2.5
Países Bajos	5.2
Japón	10
Italia	11
Canadá	12.8

- (a) Estudia la concentración de la distribución gráfica y analíticamente. Relaciona los resultados obtenidos.
- (b) Si todos los clientes rebajasen sus importaciones un 10 %, ¿cuál sería entonces el grado de concentración?
13. La recaudación, en miles de pesetas, correspondiente a cierta película en las salas de cine españolas, se distribuye, por provincias, como se indica en la tabla siguiente:

Recaudación	Número de provincias
0-500	42
500-700	5
700-1500	3
4712	1
5630	1

- (a) Construir la curva de Lorenz. Interpretación.
- (b) Construir el índice de Gini. Interpretación.
14. En la tabla siguiente aparecen recogidos la distribución de los salarios anuales, en cientos de miles de unidades monetarias, de la empresa *A* y los valores p_i y q_i para la distribución de los salarios anuales de la empresa *B*:

Empresa A		Empresa B		
Salario (10^5 u.m.)	Número empleados	Salario (10^5 u.m.)	p_i	q_i
20 – 30	20	20 – 30	42,9	25
30 – 40	10	30 – 40	57,1	36,7
40 – 50	10	40 – 60	71,4	53,3
50 – 60	20	60 – 80	100	100

- (a) ¿En cuál de las dos empresas la masa salarial está distribuída más equitativamente entre los empleados?

- (b) ¿Qué porcentaje del total de la masa salarial de la empresa A acumulan el 50 % de los empleados que menos salario perciben?
- (c) Representa gráficamente las curvas de Lorenz de ambas distribuciones.

Recursos

Seguidamente exponemos una serie de páginas de la web donde podemos encontrar algunos recursos electrónicos relacionados con esta parte:

- (a) [http : //www.ruf.rice.edu/ ~ lane/stat_sim/descriptive/index.html](http://www.ruf.rice.edu/~lane/stat_sim/descriptive/index.html) de la Universidad de Rice (Texas). Permite ir modificando la forma del histograma mediante el ratón y automáticamente van cambiando las medidas de síntesis numérica de datos asociados al mismo.
- (b) [http : //www.stat.sc.edu/ ~ west/javahtml/Histogram.html](http://www.stat.sc.edu/~west/javahtml/Histogram.html) del Departamento de Estadística de la Universidad de Carolina del Sur, muestra como varía la forma de un histograma al variar la anchura de los intervalos.
- (c) [http : //www.kuleuven.ace.be/ucs/java/index.htm](http://www.kuleuven.ace.be/ucs/java/index.htm) de la Universidad de Leuven (Bélgica). Se va viendo como evoluciona el histograma y a la vez el diagrama de caja y bigotes de los datos. Al final se pueden comparar las formas de ambos gráficos.

Parte II

**DISTRIBUCIONES
BIDIMENSIONALES**

Capítulo 6

Variables estadísticas bidimensionales

6.1. Distribución conjunta

En temas precedentes hemos realizado el estudio de las distribuciones unidimensionales, obtenidas al estudiar una determinada característica sobre los elementos de una población. En el presente tema vamos a introducir las distribuciones bidimensionales, que surgen cuando analizamos simultáneamente dos características sobre cada elemento de la población.

EJEMPLO 6.1 *Consideremos el estudio, realizado sobre un conjunto de familias, de las siguientes características $X = \text{"Consumo mensual en miles de pesetas"}$ e $Y = \text{"Renta mensual en miles de pesetas"}$.*

$X \backslash Y$	50 – 100	100 – 150	150 – 200	200 – 250	$n_{i\cdot}$
0 – 50	14	5	2	0	21
50 – 150	2	19	62	4	87
150 – 300	0	0	1	65	66
300 – 350	0	0	0	26	26
$n_{\cdot j}$	16	24	65	95	$N = 200$

Supongamos que tenemos una población cuyos elementos son clasificados según dos características cuantitativas, que llamaremos X e Y . Sus diferentes valores los representaremos por x_i e y_j , respectivamente, con $i = 1, 2, \dots, k$ y $j = 1, 2, \dots, h$.

Definición 6.1 *Se denomina distribución bidimensional de frecuencias al conjunto de valores $\{(x_i, y_j; n_{ij})\}$, donde n_{ij} es la frecuencia absoluta conjunta del*

par (x_i, y_j) y $N = \sum_{i=1}^k \sum_{j=1}^h n_{ij}$ es la frecuencia total.

y_j	y_1	y_2	$\cdots$	y_j	$\cdots$	y_h	$n_{i\cdot}$
x_i							
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1h}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2h}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ih}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kh}	$n_{k\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot h}$	N

donde $n_{i\cdot} = \sum_{j=1}^h n_{ij}$; $n_{\cdot j} = \sum_{i=1}^k n_{ij}$; $N = \sum_{i=1}^k n_{i\cdot} = \sum_{j=1}^h n_{\cdot j}$.

A la anterior tabla de doble entrada se la conoce como *tabla de correlación*.

OBSERVACIÓN 6.1

- (a) Si dividimos cada frecuencia de la tabla anterior entre el número total de elementos observados, N , obtendríamos una nueva tabla, semejante a la primera, salvo que reflejaría las proporciones o frecuencias relativas. Llamaremos f_{ij} a la frecuencia relativa conjunta del par (x_i, y_j) , $f_{ij} = \frac{n_{ij}}{N}$. Análogamente se definen las frecuencias relativas marginales

$$f_{i\cdot} = \frac{n_{i\cdot}}{N} = \sum_{j=1}^h f_{ij} \quad y \quad f_{\cdot j} = \frac{n_{\cdot j}}{N} = \sum_{i=1}^k f_{ij},$$

siendo

$$\sum_{i=1}^k \sum_{j=1}^h f_{ij} = \sum_{i=1}^k f_{i\cdot} = \sum_{j=1}^h f_{\cdot j} = 1$$

- (b) Como en el caso unidimensional, las distribuciones bidimensionales pueden venir expresadas con valores de la variable agrupados en intervalos o sin agrupar. También puede ocurrir que las características en estudio tengan distinta naturaleza.

Otra forma de disponer los resultados, a veces más cómoda cuando no hay muchos pares de valores distintos, es

x_i	y_j	n_{ij}
x_1	y_1	n_{11}
$\vdots$	$\vdots$	$\vdots$
x_i	y_j	n_{ij}
$\vdots$	$\vdots$	$\vdots$
x_k	y_h	n_{kh}

EJEMPLO 6.2 Se realiza el estudio conjunto de las características $X = \text{"Edad"}$ e $Y = \text{"Salario en miles de pesetas de 100 personas"}$.

x_i	$l_{j-1} - l_j$	n_{ij}
20	50 – 100	10
20	100 – 150	3
20	150 – 200	2
21	50 – 100	5
21	100 – 150	15
21	150 – 200	5
22	50 – 100	2
22	100 – 150	20
22	150 – 200	15
23	100 – 150	13
23	150 – 200	10
		$N = 100$

Que puesta en forma de tabla de correlación, quedaría:

x_i	$l_{j-1} - l_j$	50 – 100	100 – 150	150 – 200	$n_{i.}$
20		10	3	2	15
21		5	15	5	25
22		2	20	15	37
23		0	13	10	23
$n_{.j}$		17	51	32	$N = 100$

OBSERVACIÓN 6.2 Si se realiza el estudio conjunto de dos características cualitativas, la tabla de doble entrada que resume dicha distribución bidimensional se llama *tabla de contingencia*.

EJEMPLO 6.3 Se considera el estudio simultáneo de los atributos *Sexo* y *Estado civil* de 23 individuos.

<i>E.Civil</i> <i>Sexo</i>	<i>Soltero</i>	<i>Casado</i>	<i>Viudo</i>	<i>Divorciado</i>	$n_{i\cdot}$
<i>Mujer</i>	5	4	1	0	10
<i>Hombre</i>	4	7	0	2	13
$n_{\cdot j}$	9	11	1	2	$N = 23$

6.2. Distribuciones marginales y condicionadas

A veces interesa estudiar aisladamente cada una de las variables. De esta forma obtendríamos dos distribuciones unidimensionales, que serían las correspondientes a cada una de las variables X e Y . A estas distribuciones se les llama distribuciones marginales.

Definición 6.2 *La distribución marginal de X es la distribución que sigue la variable X independientemente de los valores de la variable Y .*

x_i	$n_{i\cdot}$	$f_{i\cdot}$
x_1	$n_{1\cdot}$	$f_{1\cdot}$
x_2	$n_{2\cdot}$	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$
x_i	$n_{i\cdot}$	$f_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$
x_k	$n_{k\cdot}$	$f_{k\cdot}$
	N	1

donde $n_{i\cdot} = \sum_{j=1}^h n_{ij}$ y $f_{i\cdot} = \frac{n_{i\cdot}}{N}$ son, respectivamente, las frecuencias absolutas

y relativas marginales de la variable X , con $i = 1, 2, \dots, k$.

Análogamente se define la distribución marginal de Y .

y_j	$n_{\cdot j}$	$f_{\cdot j}$
y_1	$n_{\cdot 1}$	$f_{\cdot 1}$
y_2	$n_{\cdot 2}$	$f_{\cdot 2}$
$\vdots$	$\vdots$	$\vdots$
y_j	$n_{\cdot j}$	$f_{\cdot j}$
$\vdots$	$\vdots$	$\vdots$
y_h	$n_{\cdot h}$	$f_{\cdot h}$
	N	1

con $n_{\cdot j} = \sum_{i=1}^k n_{ij}$ y $f_{\cdot j} = \frac{n_{\cdot j}}{N}$ son, respectivamente, las frecuencias absolutas y relativas marginales de la variable Y , con $j = 1, 2, \dots, h$.

EJEMPLO 6.4 *Vamos a estudiar las distribuciones marginales que se obtienen a partir de la distribución bidimensional dada por el Ejemplo 6.1, donde se clasificaban 200 familias según el consumo y la renta mensuales, medidos ambos, en miles de pesetas:*

(a) Distribución marginal de X (Consumo):

$l_{i-1} - l_i$	0 – 50	50 – 150	150 – 300	300 – 350	
$n_{i\cdot}$	21	87	66	26	$N = 200$
$f_{i\cdot}$	0,105	0,435	0,33	0,13	1

Esta marginal nos indica como se distribuyen las familias según su consumo mensual sin tener en cuenta la renta. Podemos observar, por ejemplo, que el consumo más frecuente es el comprendido entre 50.000 y 150.000 pesetas, que corresponde a un 43,5 % de las familias estudiadas.

(b) Distribución marginal de Y (Renta):

$l_{j-1} - l_j$	50 – 100	100 – 150	150 – 200	200 – 250	
$n_{\cdot j}$	16	24	65	95	$N = 200$
$f_{\cdot j}$	0,08	0,12	0,325	0,475	1

Otro tipo de distribuciones unidimensionales que se obtienen a partir de las bidimensionales son las distribuciones condicionadas. Son distribuciones que se obtienen manteniendo fijo un valor en una de las variables y considerando los valores que toma la otra con sus respectivas frecuencias.

Definición 6.3 *La distribución condicionada de X respecto de $Y = y_j$ es la distribución que sigue la variable X cuando la variable Y toma el valor y_j .*

Si escribimos las frecuencias condicionadas absolutas como $n_{i/j}$ y las frecuencias condicionadas relativas como $f_{i/j} = \frac{n_{ij}}{n_{\cdot j}}$ (proporción de valores, entre los que $Y = y_j$, para los cuales $X = x_i$, con $i = 1, 2, \dots, k$) entonces la distribución

condicionada de X respecto de $Y = y_j$ vendrá dada por:

$x_i/Y = y_j$	$n_{i/j}$	$f_{i/j}$
x_1	n_{1j}	$f_{1/j}$
x_2	n_{2j}	$f_{2/j}$
$\vdots$	$\vdots$	$\vdots$
x_i	n_{ij}	$f_{i/j}$
$\vdots$	$\vdots$	$\vdots$
x_k	n_{kj}	$f_{k/j}$
	$n_{.j}$	1

De forma análoga se obtiene la distribución condicionada de Y respecto de $X = x_i$, distribución de los valores de Y cuando X toma el valor x_i .

$y_j/X = x_i$	$n_{j/i}$	$f_{j/i}$
y_1	n_{i1}	$f_{1/i}$
y_2	n_{i2}	$f_{2/i}$
$\vdots$	$\vdots$	$\vdots$
y_j	n_{ij}	$f_{j/i}$
$\vdots$	$\vdots$	$\vdots$
y_h	n_{ih}	$f_{h/i}$
	$n_{i.}$	1

con $n_{j/i}$ las frecuencias condicionadas absolutas y $f_{j/i} = \frac{n_{ij}}{n_{i.}}$ las frecuencias condicionadas relativas (proporción de valores, entre los que $X = x_i$, para los cuales $Y = y_j$, con $j = 1, 2, \dots, h$).

EJEMPLO 6.5 Consideremos de nuevo el enunciado del Ejemplo 6.1. Estudiemos las distribuciones condicionadas del consumo (X) respecto de una renta (Y) entre 200000 y 250000 pesetas; y de la renta respecto de un consumo entre 50000 y 150000 pesetas.

(a) Distribución condicionada de X respecto de $Y = y_4$ ($200 < Y \leq 250$).

$x_i/Y = y_4$	$n_{i/4}$	$f_{i/4}$
0 – 50	0	0
50 – 150	4	0,04
150 – 300	65	0,68
300 – 350	26	0,28
	95	1

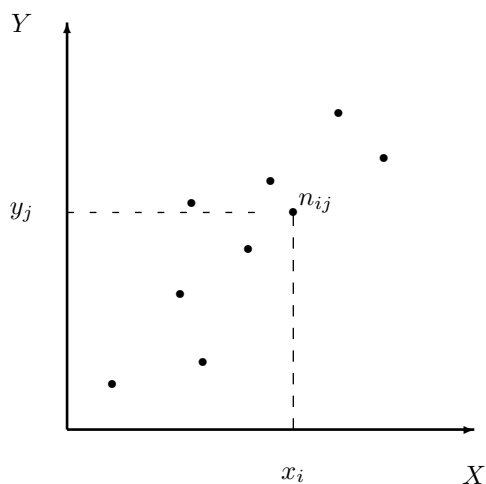
(b) *Distribución condicionada de Y respecto de $X = x_2$ ($50 < X \leq 150$).*

$y_j/X = x_2$	$n_{j/2}$	$f_{j/2}$
50 – 100	2	0,023
100 – 150	19	0,218
150 – 200	62	0,713
200 – 250	4	0,046
	87	1

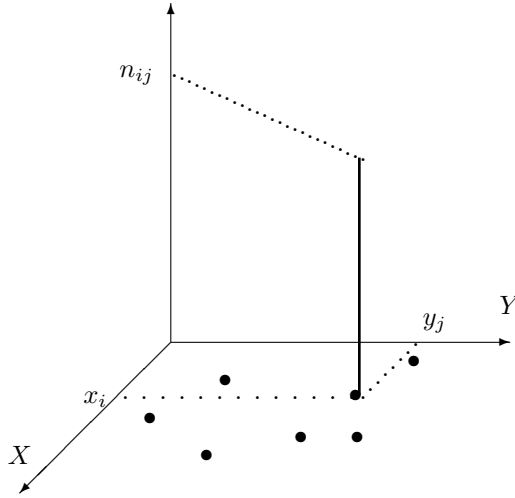
6.3. Representaciones gráficas

La representación gráfica de la distribución de frecuencias de una variable bidimensional (X, Y) , varía según la naturaleza de las variables que la componen.

(a) *Diagrama de dispersión o nube de puntos:* Consiste en representar los pares (x_i, y_j) sobre un sistema de ejes cartesianos.

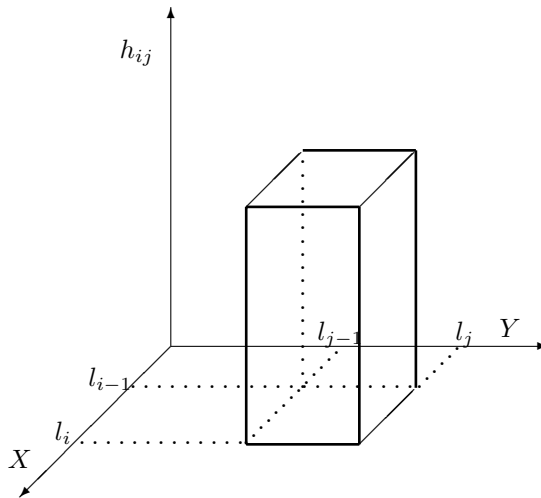


(b) *Diagrama de barras:* cada barra tendrá altura igual a la correspondiente frecuencia absoluta conjunta (o relativa).



- (c) Histograma: para el caso en que ambas variables sean continuas. Recibe también el nombre de *estereograma* y se construye levantando en cada región del plano $(l_{i-1}, l_i] \times (l_{j-1}, l_j]$ un prisma de volumen proporcional a la frecuencia absoluta conjunta.

$$V_{ij} = n_{ij} = (l_i - l_{i-1})(l_j - l_{j-1})h_{ij} \implies h_{ij} = \frac{n_{ij}}{(l_i - l_{i-1})(l_j - l_{j-1})}$$



6.4. Momentos no centrados y centrados

Tratamos de construir unos parámetros que resuman y pongan de manifiesto determinados aspectos suyos.

Definición 6.4 *Se define el momento con respecto a origen o momento ordinario o momento no centrado de orden r y s , para la distribución bidimensional $\{(x_i, y_j; n_{ij})\}$ como*

$$a_{rs} = \sum_{i=1}^k \sum_{j=1}^h x_i^r y_j^s \cdot \frac{n_{ij}}{N}$$

Casos particulares:

(1)

$$a_{00} = \sum_{i=1}^k \sum_{j=1}^h x_i^0 y_j^0 \cdot \frac{n_{ij}}{N} = 1$$

(2) Media de la distribución marginal de X .

$$a_{10} = \sum_{i=1}^k \sum_{j=1}^h x_i^1 y_j^0 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^k x_i \sum_{j=1}^h \frac{n_{ij}}{N} = \sum_{i=1}^k x_i \cdot \frac{n_{i\cdot}}{N} = \bar{x}$$

(3) Media de la distribución marginal de Y .

$$a_{01} = \sum_{i=1}^k \sum_{j=1}^h x_i^0 y_j^1 \cdot \frac{n_{ij}}{N} = \sum_{j=1}^h y_j \sum_{i=1}^k \frac{n_{ij}}{N} = \sum_{j=1}^h y_j \cdot \frac{n_{\cdot j}}{N} = \bar{y}$$

(4)

$$a_{11} = \sum_{i=1}^k \sum_{j=1}^h x_i y_j \cdot \frac{n_{ij}}{N}$$

(5) Momento respecto al origen de orden dos para la distribución marginal de X .

$$a_{20} = \sum_{i=1}^k \sum_{j=1}^h x_i^2 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^k x_i^2 \sum_{j=1}^h \frac{n_{ij}}{N} = \sum_{i=1}^k x_i^2 \cdot \frac{n_{i\cdot}}{N}$$

(6) Momento respecto al origen de orden dos para la distribución marginal de Y .

$$a_{02} = \sum_{i=1}^k \sum_{j=1}^h y_j^2 \cdot \frac{n_{ij}}{N} = \sum_{j=1}^h y_j^2 \sum_{i=1}^k \frac{n_{ij}}{N} = \sum_{j=1}^h y_j^2 \cdot \frac{n_{\cdot j}}{N}$$

Definición 6.5 Se define el momento con respecto a las medias o momento centrado de orden r y s , para la distribución bidimensional $\{(x_i, y_j; n_{ij})\}$ como

$$m_{rs} = \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})^r (y_j - \bar{y})^s \cdot \frac{n_{ij}}{N}$$

Casos particulares:

(1)

$$m_{00} = \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})^0 (y_j - \bar{y})^0 \cdot \frac{n_{ij}}{N} = 1$$

(2)

$$m_{10} = \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x}) \cdot \frac{n_{ij}}{N} = \sum_{i=1}^k (x_i - \bar{x}) \cdot \frac{n_{i\cdot}}{N} = 0$$

(3)

$$m_{01} = \sum_{i=1}^k \sum_{j=1}^h (y_j - \bar{y}) \cdot \frac{n_{ij}}{N} = \sum_{j=1}^h (y_j - \bar{y}) \cdot \frac{n_{\cdot j}}{N} = 0$$

(4) Varianza de la distribución marginal de X .

$$m_{20} = \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})^2 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^k (x_i - \bar{x})^2 \sum_{j=1}^h \frac{n_{ij}}{N} = \sum_{i=1}^k (x_i - \bar{x})^2 \cdot \frac{n_{i\cdot}}{N} = s_X^2$$

(5) Varianza de la distribución marginal de Y .

$$m_{02} = \sum_{i=1}^k \sum_{j=1}^h (y_j - \bar{y})^2 \cdot \frac{n_{ij}}{N} = \sum_{j=1}^h (y_j - \bar{y})^2 \sum_{i=1}^k \frac{n_{ij}}{N} = \sum_{j=1}^h (y_j - \bar{y})^2 \cdot \frac{n_{\cdot j}}{N} = s_Y^2$$

(6)

$$m_{11} = \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})(y_j - \bar{y}) \cdot \frac{n_{ij}}{N}$$

OBSERVACIÓN 6.3 De igual forma que los momentos de las distribuciones unidimensionales, los momentos centrados de una distribución bidimensional pueden expresarse en función de los momentos no centrados o respecto del origen:

$$m_{20} = a_{20} - a_{10}^2 = s_X^2$$

$$m_{02} = a_{02} - a_{01}^2 = s_Y^2$$

$$\begin{aligned}
m_{11} &= \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})(y_j - \bar{y}) \cdot \frac{n_{ij}}{N} = \sum_{i=1}^k \sum_{j=1}^h (x_i y_j - \bar{x} y_j - x_i \bar{y} + \bar{x} \cdot \bar{y}) \cdot \frac{n_{ij}}{N} = \\
&= \sum_{i=1}^k \sum_{j=1}^h x_i y_j - \bar{x} \cdot \sum_{i=1}^k \sum_{j=1}^h y_j \cdot \frac{n_{ij}}{N} - \bar{y} \cdot \sum_{i=1}^k \sum_{j=1}^h x_i \cdot \frac{n_{ij}}{N} + \bar{x} \cdot \bar{y} \cdot \sum_{i=1}^k \sum_{j=1}^h \frac{n_{ij}}{N} = \\
&= \sum_{i=1}^k \sum_{j=1}^h x_i y_j - \bar{x} \cdot \sum_{j=1}^h y_j \cdot \frac{n_{.j}}{N} - \bar{y} \cdot \sum_{i=1}^k x_i \cdot \frac{n_{i.}}{N} + \bar{x} \cdot \bar{y} = a_{11} - a_{10}a_{01}
\end{aligned}$$

6.5. Independencia de variables estadísticas

Definición 6.6 Diremos que las variables X e Y son estadísticamente independientes si para todo (i, j) se tiene $f_{i/j} = f_{i.}$ y $f_{j/i} = f_{.j}$.

O equivalentemente:

Dos variables X e Y son estadísticamente independientes si para todo (i, j) se verifica que

$$f_{ij} = f_{i.} f_{.j}$$

OBSERVACIÓN 6.4 Si la igualdad anterior no se verifica para algún par (i, j) , diremos que las variables son dependientes.

EJEMPLO 6.6 Estúdiese si se pueden considerar independientes las variables X e Y dadas por:

(a)

x_i/Y	y_1	y_2	y_3	y_4	$n_{i.}$
x_1	3	5	2	4	14
x_2	6	10	4	8	28
x_3	12	20	8	16	56
$n_{.j}$	21	35	14	28	$N = 98$

(b)

x_i/Y	y_1	y_2	y_3	$n_{i.}$
x_1	12	0	4	16
x_2	0	7	0	7
$n_{.j}$	12	7	4	$N = 23$

(a) X e Y son independientes pues $f_{ij} = f_{i.} f_{.j}$ para todo (i, j) .

(b) Como $f_{11} = \frac{12}{23} \neq \frac{16}{23} \frac{12}{23}$, entonces X e Y no pueden ser independientes.

6.6. Dependencia de variables estadísticas

Dos variables pueden estar relacionadas de diversas maneras:

Definición 6.7

- (a) Diremos que X e Y dependen funcionalmente si podemos establecer una aplicación que nos transforme los valores de una de las variables en los de la otra.
- (b) Diremos que X e Y dependen estadísticamente cuando la variación de una de las variables influye en la distribución de la otra. En este caso es imposible encontrar una ecuación tal que los valores observados para dichas variables la satisfagan. Gráficamente, equivale al hecho de que no es posible encontrar una función tal que su gráfica pase por todos los puntos correspondientes al diagrama de dispersión asociado a las variables observadas. Dentro de la dependencia estadística nos podemos encontrar a su vez con:
- (i) Dependencia lineal, es decir, las dos variables admiten una forma funcional en forma de una línea recta.
- (ii) Otro tipo de dependencia: hiperbólica, exponencial, potencial, etc.

EJEMPLO 6.7 La siguiente tabla recoge todas las distribuciones condicionadas relativas de X (consumo mensual) respecto de Y (renta mensual) correspondientes al Ejemplo 6.1.

$l_{i-1} - l_i/Y$	$f_{i/1}$	$f_{i/2}$	$f_{i/3}$	$f_{i/4}$
0 – 50	0,8754	0,208	0,031	0
50 – 150	0,125	0,792	0,954	0,042
150 – 300	0	0	0,015	0,684
300 – 350	0	0	0	0,274
	1	1	1	1

Podemos observar el diferente comportamiento del consumo según los diferentes valores de la renta. A medida que consideramos valores más altos de la renta observamos porcentajes mayores de familias con consumos más elevados. Por tanto, podemos pensar que Consumo y Renta deben estar relacionadas.

6.7. Dependencia lineal. Covarianza

Definición 6.8 Se define la covarianza de las variables X e Y como el momento centrado m_{11} . La representaremos por s_{XY} . Es decir:

$$s_{XY} = m_{11} = \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{X})(y_j - \bar{Y}) \cdot \frac{n_{ij}}{N}$$

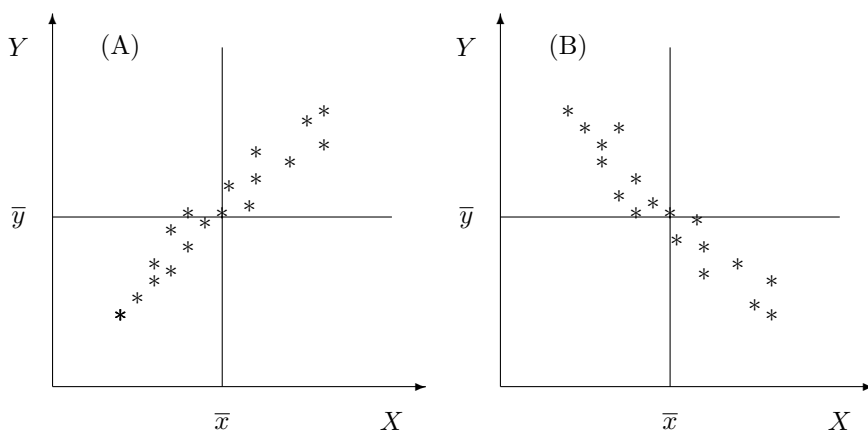
OBSERVACIÓN 6.5

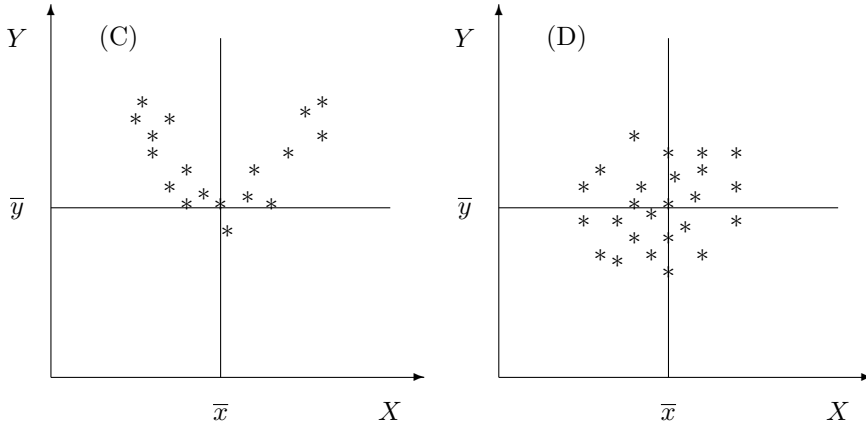
- (a) Generalmente la covarianza se calcula, teniendo en cuenta la observación 6.3 mediante la expresión siguiente:

$$s_{XY} = a_{11} - a_{10}a_{01}$$

- (b) La covarianza, s_{XY} , es una medida de la asociación lineal entre dos variables, que resume la información existente en un gráfico de dispersión. El empleo de las variables respecto de la media equivale a una traslación de los ejes de coordenadas. El centro de coordenadas estará ahora en el centro de la nube de puntos. Si la relación es positiva los puntos aparecerán mayoritariamente en los cuadrantes primero y tercero, y si es negativa en los cuadrantes segundo y cuarto.
- (c) La covarianza como medida de asociación lineal, cambia si modificamos las unidades de medida de las variables. Esto no nos permite comparar pares de variables medidas en unidades diferentes.

EJEMPLO 6.8 Observamos estas nubes de puntos:





Cuando X e Y varían conjuntamente de forma lineal, gráficas (A) y (B), la covarianza será alta. Cuando no exista relación entre X e Y , gráfica (D), o exista una relación marcadamente no lineal, gráfica (C), la covarianza será cero.

Si $s_{XY} > 0 \Rightarrow X$ e Y varían de forma lineal en el mismo sentido que llamaremos asociación lineal directa o positiva, gráfica (A).

Si $s_{XY} < 0 \Rightarrow X$ e Y varían de forma lineal en sentido opuesto o diremos que presentan asociación lineal inversa o negativa, gráfica (B).

Definición 6.9 Diremos que las variables X e Y son incorreladas si $s_{XY} = 0$ (ausencia de asociación lineal).

OBSERVACIÓN 6.6

(a) Si X e Y son independientes $\Rightarrow X$ e Y son incorreladas.

X e Y independientes $\Rightarrow$

$$a_{11} = \sum_{i=1}^k \sum_{j=1}^h x_i y_j \cdot \frac{n_{ij}}{N} = \sum_{i=1}^k \sum_{j=1}^h x_i y_j \cdot \frac{n_{i\cdot}}{N} \cdot \frac{n_{\cdot j}}{N} = a_{10} a_{01}$$

$$\Rightarrow m_{11} = a_{11} - a_{10} a_{01} = 0 \Rightarrow X \text{ e } Y \text{ son incorreladas.}$$

(b) Podemos encontrar variables que sean incorreladas y no sean independientes. Si existe una relación no lineal entre las variables, ésta no tiene por qué detectarse con la covarianza.

EJEMPLO 6.9 *Consideremos la distribución bidimensional:*

x_i	y_j	-2	0	2	$n_{i.}$	$x_i n_{i.}$	$\sum_{j=1}^h x_i y_j n_{ij}$
-1		0	3	0	3	-3	$0 + 0 + 0 = 0$
1		2	9	2	13	13	$-4 + 0 + 4 = 0$
3		0	4	0	4	12	$0 + 0 + 0 = 0$
	$n_{.j}$	2	16	2	$N = 20$	22	0
	$y_j n_{.j}$	-4	0	4	0		

$$\bar{x} = a_{10} = \sum_{i=1}^k x_i \cdot \frac{n_{i.}}{N} = \frac{22}{20} = 1,1$$

$$\bar{y} = a_{01} = \sum_{j=1}^h y_j \cdot \frac{n_{.j}}{N} = \frac{0}{20} = 0$$

$$s_{XY} = a_{11} - a_{10}a_{01} = \sum_{i=1}^k \sum_{j=1}^h x_i y_j \cdot \frac{n_{ij}}{N} - \bar{x} \cdot \bar{y} = 0 - 1,1 \times 0 = 0$$

Entonces X e Y son incorreladas.

X e Y no son independientes ya que:

$$f_{11} = \frac{0}{20} \neq \frac{3}{20} \frac{2}{20} = f_{i.} f_{.j}$$

Capítulo 7

Ajustes

7.1. Introducción

Sea una distribución bidimensional $\{(x_i, y_j; n_{ij})\}$ en la que se supone que existe una relación estadística entre las variables X e Y .

Definición 7.1 *Ante la imposibilidad de encontrar una gráfica que pase por todos los puntos, el problema del ajuste consiste en buscar una función cuya gráfica se "adapte" o se "acerque" lo más posible a los datos observados, es decir, que su aproximación a los mismos sea óptima de acuerdo a un criterio adoptado, y de esta manera exprese mejor la relación entre los mismos.*

En todo ajuste se plantean dos problemas:

- (a) El primero es la elección de una familia de funciones. Ello lo haremos a través de:
 - (i) Un método gráfico, es decir, mediante la observación de la nube de puntos.
 - (ii) El análisis de los datos también nos puede ayudar a elegir la familia de funciones.
 - Así, si las abscisas son más o menos equidistantes y las ordenadas varían en progresión aritmética, la familia adecuada es la lineal $y = a + bx$.
 - Si las ordenadas varían en progresión geométrica la familia adecuada es la exponencial $y = ab^x$.
 - O si estamos ante un fenómeno económico con dos variables con elasticidad constante, podríamos decantarnos por la familia potencial $y = ax^b$, ya que su elasticidad, $\mathcal{E}$, es

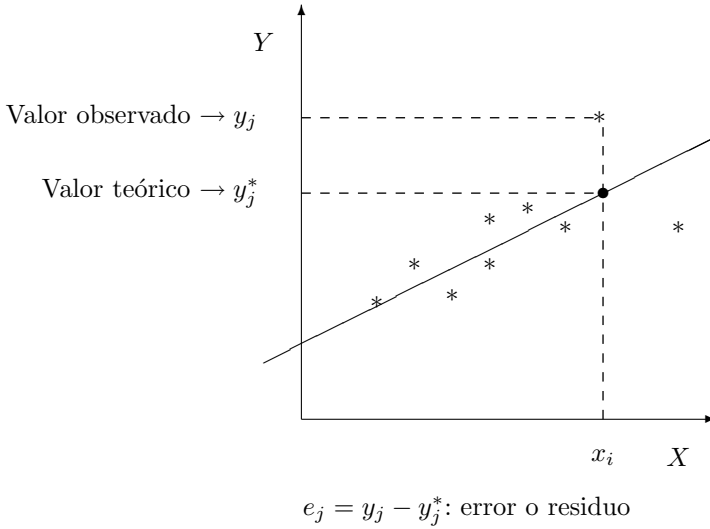
$$\mathcal{E}(y) = y' \frac{y}{y} \Rightarrow \mathcal{E}(ax^b) = abx^{b-1} \frac{x}{ax^b} = b$$

El concepto de elasticidad representa una medida de la sensibilidad que muestra una determinada variable cuando se modifica cualquier otra de la que depende. Es el cociente entre variaciones porcentuales de las dos variables y se puede calcular como la pendiente de la curva por el cociente entre las variables.

- (b) La identificación u obtención de la función de ajuste. Se realiza mediante la asignación de un conjunto de valores concretos a los parámetros desconocidos, aplicando técnicas matemáticas de optimización, acordes con el criterio óptimo establecido.

7.1.1. El método de mínimos cuadrados

Una vez elegida la familia a la que pertenece la curva, $Y = f(X; a_1, \dots, a_n)$,



Definición 7.2 El método de mínimos cuadrados consiste en calcular los parámetros de la curva seleccionada utilizando como criterio "la minimización de la suma de los cuadrados de los residuos o errores".

El procedimiento a seguir es el siguiente:

- (a) Primer paso: elegir la función $Y = f(X; a_1, \dots, a_n)$
- (b) Segundo paso: plantear

$$\min \sum_{i=1}^k \sum_{j=1}^h (y_j - y_j^*)^2 n_{ij} = \min \sum_{i=1}^k \sum_{j=1}^h (y_j - f(x_i; a_1, \dots, a_n))^2 n_{ij} = \min \Phi$$

(c) Tercer paso: resolver el "sistema de ecuaciones normales"

$$s.e.n. \hookrightarrow \begin{cases} \frac{\partial \Phi}{\partial a_1} = 0 \\ \frac{\partial \Phi}{\partial a_2} = 0 \\ \vdots \\ \frac{\partial \Phi}{\partial a_n} = 0 \end{cases}$$

OBSERVACIÓN 7.1 Otros posibles criterios para realizar el ajuste hubieran sido:

(a) Minimizar la suma de residuos, esto es,

$$\min \sum_{i=1}^k \sum_{j=1}^h (y_j - y_j^*) n_{ij}$$

pero este criterio no es adecuado, ya que serían igualmente buenos ajustes con suma de residuos nula, conseguidos con residuos nulos o compensando residuos positivos o negativos.

(b) Minimizar la suma del valor absoluto de los residuos, esto es,

$$\min \sum_{i=1}^k \sum_{j=1}^h |y_j - y_j^*| n_{ij}$$

tampoco este criterio es el adecuado por no prestarse a manipulaciones algebraicas sencillas.

7.2. Ajuste lineal. Función de consumo de Keynes

(a) Primer paso: elegimos la función $Y = f(X; a, b) = a + bX$

(b) Segundo paso: planteamos

$$\min \sum_{i=1}^k \sum_{j=1}^h (y_j - y_j^*)^2 n_{ij} = \min \sum_{i=1}^k \sum_{j=1}^h (y_j - a - bx_i)^2 n_{ij} = \min \Phi$$

(c) Tercer paso: resolvemos el "sistema de ecuaciones normales"

$$s.e.n. \hookrightarrow \begin{cases} \frac{\partial \Phi}{\partial a} = -2 \sum_{i=1}^k \sum_{j=1}^h (y_j - a - bx_i) n_{ij} = 0 \\ \frac{\partial \Phi}{\partial b} = -2 \sum_{i=1}^k \sum_{j=1}^h (y_j - a - bx_i) x_i n_{ij} = 0 \end{cases}$$

o sea, el sistema

$$\left\{ \begin{array}{l} \sum_{j=1}^h y_j n_{.j} = aN + b \sum_{i=1}^k x_i n_{i.} \\ \sum_{i=1}^k \sum_{j=1}^h x_i y_j n_{ij} = a \sum_{i=1}^k x_i n_{i.} + b \sum_{i=1}^k x_i^2 n_{i.} \end{array} \right.$$

que, dividiendo por N se transforma en

$$\left. \begin{array}{l} \bar{y} = a + b\bar{x} \\ a_{11} = a\bar{x} + ba_{20} \end{array} \right\}$$

cuya solución es:

$$b = \frac{a_{11} - \bar{x} \cdot \bar{y}}{a_{20} - \bar{x}^2} = \frac{s_{XY}}{s_X^2}$$

$$a = \bar{y} - b\bar{x} = \bar{y} - \frac{s_{XY}}{s_X^2} \cdot \bar{x}$$

OBSERVACIÓN 7.2 *El ajuste lineal conlleva una serie de propiedades:*

- (a) *La primera ecuación del sistema de ecuaciones normales nos dice que la suma de residuos es nula y por tanto la media del error es cero. Como consecuencia de lo anterior la media observada coincide con la media de los valores teóricos.*
- (b) *La segunda ecuación normal nos dice que la suma de los productos de los residuos por los correspondientes valores de la variable independiente es nula.*
- (c) *Las dos propiedades anteriores nos permitirán, mas tarde, expresar la varianza de la variable dependiente como suma de la varianza de los valores calculados y la varianza de los residuos.*

EJEMPLO 7.1 *Las variables $Y = \text{"Consumo"}$ y $X = \text{"Renta"}$, se relacionan linealmente mediante la conocida como función Keynesiana de Consumo,*

$$Y = a + bX,$$

$x_i(\text{en miles})$	y_j	$x_i y_j$	x_i^2
20	100	2000	400
25	150	3750	625
35	180	6300	1225
40	200	8000	1600
45	210	9450	2025
$\sum_{i=1}^k x_i n_{i.}$	$\sum_{j=1}^h y_j n_{.j}$	$\sum_{i=1}^k \sum_{j=1}^h x_i y_j n_{ij}$	$\sum_{i=1}^k x_i^2 n_{i.}$
165	840	29500	5875

$N = 5$, el sistema de ecuaciones normales es:

$$\left. \begin{aligned} 840 &= 5a + 165b \\ 29500 &= 165a + 5875b \end{aligned} \right\}$$

cuyas soluciones son

$$a = \frac{1350}{43}$$

$$b = \frac{178}{43}$$

luego

$$y = \frac{1350}{43} + \frac{178}{43} \cdot x$$

Otra forma de resolverlo sería:

$$\bar{x} = \frac{\sum_{i=1}^k x_i n_i}{N} = \frac{165}{5} = 33$$

$$\bar{y} = \frac{\sum_{j=1}^h y_j n_{.j}}{N} = \frac{840}{5} = 168$$

$$s_{XY} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^h x_i y_j n_{ij} - \bar{x} \cdot \bar{y} = \frac{29500}{5} - (33)(168) = 356$$

$$s_X^2 = \frac{1}{N} \sum_{i=1}^k x_i^2 n_i - \bar{x}^2 = \frac{5875}{5} - (33)^2 = 86$$

$$b = \frac{s_{XY}}{s_X^2} = \frac{356}{86} = \frac{178}{43}$$

$$a = \bar{y} - b\bar{x} = 168 - 33 \cdot \frac{178}{43} = \frac{1350}{43}$$

$$y = \frac{1350}{43} + \frac{178}{43} \cdot x$$

7.3. Ajustes reducibles al caso lineal

Muchas veces interesa reducir ciertos ajustes al caso lineal debido a:

- (a) La sencillez del caso lineal.
- (b) Porque casi toda función puede aproximarse por una recta en un pequeño dominio.

A continuación expondremos algunos ajustes que permiten esta simplificación.

7.3.1. Ajuste hiperbólico: curvas de demanda

La función de tipo hiperbólico tiene especial interés en el ámbito económico-empresarial para relacionar la cantidad demandada de cierto bien, Y , y el precio del mismo, X .

$$Y = a + \frac{b}{X}$$

Haciendo $Z = \frac{1}{X}$, queda $Y = a + bZ$

Precios, x_i	Cantidades, y_j	$z_i = 1/x_i$	$z_i y_j$	z_i^2
1	16	1	16	1
2	8	1/2	4	1/4
4	4	1/4	1	1/16
8	2	1/8	1/4	1/64
	$\sum_{j=1}^h y_j n_{.j}$	$\sum_{i=1}^k z_i n_{i.}$	$\sum_{i=1}^k \sum_{j=1}^h z_i y_j n_{ij}$	$\sum_{i=1}^k z_i^2 n_{i.}$
	30	15/8	85/4	85/64

$N = 4$, el sistema de ecuaciones normales es:

$$s.e.n. \hookrightarrow \begin{cases} 30 = 4a + \frac{15}{8} \cdot b \\ \frac{85}{4} = \frac{15}{8} \cdot a + \frac{85}{64} \cdot b \end{cases} \Rightarrow \begin{cases} a = 0 \\ b = 16 \end{cases}$$

luego

$$y = \frac{16}{x}$$

Otra forma:

$$\begin{aligned}\bar{z} &= \frac{\sum_{i=1}^k z_i n_{i\cdot}}{N} = \frac{15}{32} \\ \bar{y} &= \frac{\sum_{j=1}^h y_j n_{\cdot j}}{N} = \frac{30}{4} = 7,5 \\ s_{ZY} &= \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^h z_i y_j n_{ij} - \bar{z} \cdot \bar{y} = \frac{85}{16} - \frac{15}{32}(7,5) = \frac{115}{64} \\ s_Z^2 &= \frac{1}{N} \sum_{i=1}^k z_i^2 n_{i\cdot} - \bar{z}^2 = \frac{85}{256} - \left(\frac{15}{32}\right)^2 = \frac{115}{1024} \\ b &= \frac{s_{ZY}}{s_Z^2} = \frac{115}{64} : \frac{115}{1024} = 16 \\ a &= \bar{y} - b\bar{z} = 7,5 - 16 \left(\frac{15}{32}\right) = 0\end{aligned}$$

$$y = \frac{16}{x}$$

7.3.2. Ajuste potencial: función de Cobb-Douglas

La función de producción de Cobb-Douglas presenta una expresión más compleja, que bajo determinadas hipótesis puede escribirse en la forma siguiente:

$$Y = aX^b$$

siendo X = "Cantidad de trabajo" e Y = "Producción". Tomando logaritmos neperianos queda $\ln Y = \ln a + b \ln X$, y haciendo $U = \ln Y$, $A = \ln a$, $Z = \ln X$ queda $U = A + bZ$

x_i	y_j	$z_i = \ln x_i$	$u_j = \ln y_j$	$z_i u_j$	z_i^2
200	10	5,298317	2,302585	12,199825	28,072163
400	20	5,991464	2,995732	17,948820	35,897640
600	26	6,396929	3,258096	20,841808	40,920700
800	31	6,684611	3,433987	22,954867	44,684024
		$\sum_{i=1}^k z_i n_{i\cdot}$	$\sum_{j=1}^h u_j n_{\cdot j}$	$\sum_{i=1}^k \sum_{j=1}^h z_i u_j n_{ij}$	$\sum_{i=1}^k z_i^2 n_{i\cdot}$
		24,371321	11,990400	73,945320	149,574527

$N = 4$, el sistema de ecuaciones normales es:

$$s.e.n. \hookrightarrow \begin{cases} 11,990400 = 4A + 24,371321b \\ 73,94532 = 24,371321A + 149,574527b \end{cases} \Rightarrow \begin{cases} A = -2,002986 \\ b = 0,820733 \end{cases}$$

Como $A = \ln a$, se tiene que $a = 0,134931$. Por tanto

$$y = 0,134931 \cdot x^{0,820733}$$

Otra forma:

$$\bar{z} = \frac{\sum_{i=1}^k z_i n_i}{N} = \frac{24,371321}{4} = 6,09283$$

$$\bar{u} = \frac{\sum_{j=1}^h u_j n_j}{N} = \frac{11,9904}{4} = 2,9976$$

$$s_{ZU} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^h z_i u_j n_{ij} - \bar{z} \cdot \bar{u} = \frac{73,94532}{4} - \frac{24,371321}{4} \cdot \frac{11,9904}{4} = 0,22246$$

$$s_Z^2 = \frac{1}{N} \sum_{i=1}^k z_i^2 n_i - \bar{z}^2 = \frac{149,574527}{4} - \left(\frac{24,371321}{4} \right)^2 = 0,271051$$

$$b = \frac{s_{ZU}}{s_Z^2} = \frac{0,222462}{0,271051} = 0,82073$$

$$A = \bar{u} - b\bar{z} = -2,0029$$

$$y = 0,134931 \cdot x^{0,820733}$$

7.3.3. Ajuste exponencial: modelo de Harrod-Domar

El modelo de crecimiento económico de Harrod-Domar relaciona la renta final, Y , con el tiempo necesario para obtenerla, X , mediante la expresión siguiente:

$$Y = y_0 e^{\rho X}$$

donde y_0 es la renta inicial y ρ es un parámetro.

Tomando logaritmos neperianos queda: $\ln Y = \ln y_0 + \rho X$ y haciendo $U = \ln Y$, $A = \ln y_0$, resulta $U = A + \rho X$

x_i	y_j	$u_j = \ln y_j$	$x_i u_j$	x_i^2
1	12	2,484906	2,484906	1
2	15	2,708050	5,416100	4
3	20	2,995732	8,987196	9
4	26	3,258096	13,032384	16
5	34	3,526360	17,631800	25
$\sum_{i=1}^k x_i n_{i.}$		$\sum_{j=1}^h u_j n_{.j}$	$\sum_{i=1}^k \sum_{j=1}^h x_i u_j n_{ij}$	$\sum_{i=1}^k x_i^2 n_{i.}$
15		14,9731144	47,552386	55

$N = 5$, el sistema de ecuaciones normales es:

$$s.e.n. \hookrightarrow \begin{cases} 14,973144 = 5A + 15\rho \\ 47,552386 = 15A + 55\rho \end{cases} \Rightarrow \begin{cases} A = 2,204742 \\ \rho = 0,263295 \end{cases}$$

Como $A = \ln y_0$, se tiene que $y_0 = 9,067911$. Luego

$$y = 9,067911 \cdot e^{0,263295x}$$

Otra forma:

$$\bar{x} = \frac{\sum_{i=1}^k x_i n_{i.}}{N} = \frac{15}{5} = 3$$

$$\bar{u} = \frac{\sum_{j=1}^h u_j n_{.j}}{N} = \frac{14,973144}{5} = 2,994628$$

$$s_{XU} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^h x_i u_j n_{ij} - \bar{x} \cdot \bar{u} = \frac{47,552386}{5} - 3 \cdot \frac{14,973144}{5} = 0,52659$$

$$s_X^2 = \frac{1}{N} \sum_{i=1}^k x_i^2 n_{i.} - \bar{x}^2 = \frac{55}{5} - \left(\frac{15}{5}\right)^2 = 2$$

$$\rho = \frac{s_{XU}}{s_X^2} = \frac{0,52659}{2} = 0,26329$$

$$A = \bar{u} - \rho \bar{x} = 2,2047 \Rightarrow y_0 = e^A = 9,067911$$

$$y = 9,067911 \cdot e^{0,263295x}$$

7.4. Otros ajustes

Existen otras curvas que presentan serios problemas para poder aplicarles directamente el método de mínimos cuadrados. Estas curvas se ajustan a los datos siguiendo otros métodos alternativos.

7.4.1. Exponencial modificada: la Ley de Makeham

Su expresión general es la siguiente:

$$Y = a + bc^X$$

Este tipo de curva no puede reducirse a otro lineal tomando logaritmos. Esta función es muy usada en Estadística Actuarial para obtener tablas de mortalidad (la Ley de Makeham que relaciona la tasa instantánea de mortalidad, Y , con la edad de la persona, X).

Para proceder a su ajuste utilizaremos el *Método de las Sumas*. Consiste básicamente en dividir la distribución en tres partes, igualar las sumas correspondientes de las ordenadas observadas y teóricas, y a partir de las igualdades resultantes se despejan los tres parámetros. A continuación detallamos el proceso a seguir:

Para ello establecemos tres grupos con igual número de datos. Si es necesario omitimos alguno de ellos, de modo que el número de observaciones sea múltiplo de tres, $n = 3h$.

Se calculan las sumas reales u observadas:

$$S_1 = \sum_{i=1}^h y_i, \quad S_2 = \sum_{i=h+1}^{2h} y_i, \quad S_3 = \sum_{i=2h+1}^{3h} y_i$$

que se igualan a las sumas correspondientes ordenadas teóricas de cada partición. Admitiendo que las abscisas varían de unidad en unidad y siendo x_1 el primer valor de la variable X , tenemos que:

$$S_1 = \sum_{i=1}^h (a + bc^{x_i}) = ha + b \sum_{i=1}^h c^i = ha + bc^{x_1} \cdot \frac{c^h - 1}{c - 1} = ha + b\lambda$$

$$S_2 = \sum_{i=h+1}^{2h} (a + bc^{x_i}) = ha + b \sum_{i=h+1}^{2h} c^i = ha + bc^{h+x_1} \cdot \frac{c^h - 1}{c - 1}$$

$$S_3 = \sum_{i=2h+1}^{3h} (a + bc^{x_i}) = ha + b \sum_{i=2h+1}^{3h} c^i = ha + bc^{2h+x_1} \cdot \frac{c^h - 1}{c - 1}$$

Luego

$$\frac{\Delta S_2}{\Delta S_1} = \frac{S_3 - S_2}{S_2 - S_1} = \frac{b \frac{c^h - 1}{c - 1} (c^{2h+x_1} - c^{h+x_1})}{b \frac{c^h - 1}{c - 1} (c^{h+x_1} - c^{x_1})} = c^h$$

$$c = \sqrt[h]{\frac{\Delta S_2}{\Delta S_1}}$$

Una vez obtenido c de la expresión ΔS_1 calcularemos b:

$$b = \frac{\Delta S_1}{c^{x_1} \cdot \frac{c^h - 1}{c - 1} (c^h - 1)} = \frac{\Delta S_1}{\lambda (c^h - 1)}$$

Y sustituidos b y c en la expresión de S_1 obtenemos el parámetro que falta

$$a = \frac{S_1 - b c^{x_1} \cdot \frac{c^h - 1}{c - 1}}{h} = \frac{S_1 - b \lambda}{h}$$

x_i	y_j	S_i
1	100	$S_1 = 190$
2	90	
3	100	
4	110	$S_2 = 210$
5	130	
6	110	$S_3 = 240$

$$c = \sqrt[h]{\frac{\Delta S_2}{\Delta S_1}} = \sqrt{\frac{240 - 210}{210 - 190}} = 1,2247$$

Una vez obtenido c de la expresión ΔS_1 calcularemos b, donde hemos llamado

$$\lambda = c^{x_1} \cdot \frac{c^h - 1}{c - 1} = 2,7251$$

$$b = \frac{\Delta S_1}{\lambda (c^h - 1)} = \frac{20}{2,7251(0,5)} = 14,6783$$

Y sustituidos b y c en la expresión de S_1 obtenemos el parámetro que falta

$$a = \frac{S_1 - b \lambda}{h} = \frac{190 - 14,6783(2,7251)}{2} = 75$$

Luego

$$y = 75 + 14,6783(1,2247)^x$$

7.4.2. La curva logística

La evolución a lo largo del tiempo de una gran cantidad de magnitudes socioeconómicas como el desarrollo del comercio exterior, la venta de nuevos productos o el crecimiento de poblaciones, etc. pueden expresarse mediante la curva logística

$$Y = \frac{a}{1 + be^{-cX}}$$

donde X suele representar el tiempo y la variable Y la magnitud en estudio.

Las características de esta curva son: es una curva creciente, con una fase inicial de crecimiento lento, una fase de rápido crecimiento en la que prácticamente se alcanza el tope, y que se corresponde con un período de expansión y mediante un punto de inflexión se pasa a una última fase de crecimiento lento correspondiente a un período de estabilización. La curva se encuentra entre dos asíntotas horizontales $y=a$ e $y=0$. Fue propuesta en 1845 por el matemático belga Pierre François Verhulst.

Uno de los métodos para ajustar una distribución a una curva logística es el llamado *Método de los Valores Recíprocos*. A grandes rasgos este método consiste en invertir la función a ajustar y aplicar el método de las sumas.

Sea

$$Y = \frac{a}{1 + be^{-cX}} \Rightarrow \frac{1}{Y} = \frac{1}{a} + \frac{b}{a}e^{-cX}$$

notamos

$$\frac{1}{Y} = Z, \quad \frac{1}{a} = A, \quad \frac{b}{a} = B, \quad e^{-c} = C$$

entonces nos queda una exponencial modificada

$$Z = A + BC^X$$

y los valores de A, B y C se pueden calcular por el método de las sumas.

x_i	y_j	$Z = \frac{1}{Y}$	S_i
1	40,3	0,0248	$S_1 = 0,0458$
2	47,6	0,0210	
3	55,1	0,0181	
4	62,4	0,0160	$S_2 = 0,0341$
5	69,1	0,0145	
6	75,1	0,0133	$S_3 = 0,0278$

$$C = \sqrt[n]{\frac{\Delta S_2}{\Delta S_1}} = \sqrt{\frac{0,0278 - 0,0341}{0,0341 - 0,0458}} = 0,7338$$

$$c = \ln\left(\frac{1}{C}\right) = 0,3095$$

Una vez obtenido C de la expresión ΔS_1 calcularemos B, donde hemos llamado

$$\lambda = C^{x_1} \cdot \frac{C^h - 1}{C - 1} = 1,2722$$

es decir,

$$B = \frac{\Delta S_1}{\lambda(C^h - 1)} = \frac{-0,0117}{1,2722(0,7338^2 - 1)} = 0,02$$

Y sustituidos B y C en la expresión de S_1 obtenemos el parámetro que falta

$$A = \frac{S_1 - B\lambda}{h} = \frac{0,0458 - 0,02(1,2722)}{2} = 0,0102$$

Y por tanto, deshaciendo el cambio $a = \frac{1}{A} = 98,04$, $b = \frac{B}{A} = 1,968$.
Luego

$$y = \frac{98,04}{1 + 1,96e^{-0,3095x}}$$

Capítulo 8

Regresión simple

8.1. Introducción al concepto de regresión

Definición 8.1 *La regresión tiene por objeto poner de manifiesto, a partir de la información de que se disponga, la estructura de dependencia que mejor explique el comportamiento de una variable Y (variable dependiente o explicada) a través de un conjunto de variables $X_1, X_2, \dots, X_p$ (variables independientes o explicativas), con las que se supone que está relacionada.*

El caso que nos disponemos a estudiar es el más sencillo, utiliza una sola variable explicativa, y se conoce como *Regresión Simple*.

Definición 8.2 *Sea una distribución bidimensional $\{(x_i, y_j; n_{ij})\}$, llamaremos regresión de Y sobre X a la función que explica la variable Y para cada valor de la X .*

La regresión de X sobre Y nos hablará del comportamiento de X para cada valor de Y .

Galton utilizó el término "*regresión*" para explicar que la estatura de los hijos tendía a "*regresar*" a la media de la población. Es decir, de padres más bajos que la media de la población, nacen hijos más altos que sus progenitores; y de padres más altos que la media nacen hijos más bajos que ellos.

8.2. Regresión de la media

Uno de los objetivos más importantes del análisis de la regresión es predecir el valor de una variable dependiente Y dado el valor x_i de una variable independiente asociada a X , es decir, predecir el comportamiento de la variable condicionada $Y/X = x_i$. Si queremos asignar un valor a la variable Y para un x_i cualquiera, podemos tomar la media condicionada de Y a ese valor x_i . Hemos optado por

sustituir todos los puntos de cada distribución condicionada por uno sólo: su media. Son bien conocidas las cualidades de la media aritmética como valor representativo del comportamiento de una variable, aunque otro criterio hubiera podido ser asociar en vez de la media la mediana.

EJEMPLO 8.1 *Consideremos la distribución bidimensional dada por la siguiente tabla:*

y_j x_i	2	3	4	$n_{i.}$
1	1	3	0	4
2	4	4	2	10
3	0	2	1	3
$n_{.j}$	5	9	3	$N = 17$

Queremos calcular la regresión de la media de Y sobre X .

$y_j/X = x_1$	$n_{j/1}$	$y_j/X = x_2$	$n_{j/2}$	$y_j/X = x_3$	$n_{j/3}$
2	1	2	4	2	0
3	3	3	4	3	2
4	0	4	2	4	1

$$(\bar{y}/X = x_1) = \frac{2(1) + 3(3) + 4(0)}{4} = \frac{11}{4} = 2,75$$

$$(\bar{y}/X = x_2) = \frac{2(4) + 3(4) + 4(2)}{10} = \frac{28}{10} = 2,8$$

$$(\bar{y}/X = x_3) = \frac{2(0) + 3(2) + 4(1)}{3} = \frac{10}{3} = 3,33$$

Los puntos que constituyen la llamada "línea de regresión" son, en este caso:

$$(1; 2,75); (2; 2,8); (3; 3,33)$$

Si queremos asignar un valor a la variable X para un valor y_j de Y , podemos tomar la media condicionada de X a ese valor y_j .

EJEMPLO 8.2 *Dada la distribución bidimensional del Ejemplo 8.1,*

y_j x_i	2	3	4	$n_{i.}$
1	1	3	0	4
2	4	4	2	10
3	0	2	1	3
$n_{.j}$	5	9	3	$N = 17$

Vamos a calcular la regresión de la media de X sobre Y .

$x_i/Y = y_1$	$n_{i/1}$	$x_i/Y = y_2$	$n_{i/2}$	$x_i/Y = y_3$	$n_{i/3}$
1	1	1	3	1	0
2	4	2	4	2	2
3	0	3	2	3	1

$$(\bar{x}/Y = y_1) = \frac{1(1) + 2(4) + 3(0)}{5} = \frac{9}{5} = 1,8$$

$$(\bar{x}/Y = y_2) = \frac{1(3) + 2(4) + 3(2)}{9} = \frac{17}{9} = 1,88$$

$$(\bar{x}/Y = y_3) = \frac{1(0) + 2(2) + 3(1)}{3} = \frac{7}{3} = 2,33$$

Los puntos que constituyen la llamada "línea de regresión" son, en este caso:

$$(1,8; 2); (1,88; 3); (2,33; 4)$$

OBSERVACIÓN 8.1

- En la regresión de la media no necesitamos fijar de antemano el tipo de función.
- Utiliza el promedio más significativo, la media.
- La curva así definida tiene la importante propiedad de ser, entre todas las funciones, la que mejor se ajusta a los datos observados según el método de mínimos cuadrados. De todas las posibles curvas de regresión mínimo-cuadrática, la óptima es la regresión de la media. Por lo tanto, si la aplicación que define la aplicación de la media tiene una expresión analítica, siempre identificará a la curva óptima de regresión mínimo-cuadrática.
- Este procedimiento presenta el inconveniente de que, en general, trabajamos con un número finito de puntos aislados, a pesar de ello la denominamos curva, y ello hace que este tipo de regresión sea poco manejable para conseguir nuestro objetivo: explicar una variable a través del comportamiento de la otra y efectuar predicciones.
- Es un procedimiento muy laborioso.

8.3. Regresión mínimo-cuadrática

En este caso tomaremos como valor de una variable, para uno dado de la otra, el deducido de la función ajustada por mínimos cuadrados.

(a) Regresión de Y sobre X

$$\min \sum_{i=1}^k \sum_{j=1}^h (y_j - y_j^*)^2 n_{ij}$$

(b) Regresión de X sobre Y

$$\min \sum_{i=1}^k \sum_{j=1}^h (x_i - x_i^*)^2 n_{ij}$$

con $y_j^* = f(x_i)$ y $x_i^* = g(y_j)$, siendo f y g las funciones seleccionadas mediante el estudio de la nube de puntos correspondiente.

OBSERVACIÓN 8.2

- (a) *Con la regresión mínimo-cuadrática, la elección de la función se debe hacer por anticipado.*
- (b) *La regresión mínimo-cuadrática es tanto mejor cuanto más acertada sea la elección de la curva.*
- (c) *La regresión a la media coincide con la regresión mínimo cuadrática si la media de la distribución condicionada es una función lineal.*
- (d) *La regresión mínimo-cuadrática es un procedimiento sencillo y que permite hacer predicciones.*

8.3.1. Regresión lineal mínimo-cuadrática

(a) Regresión de Y sobre X

$$y_j^* = a + bx_i$$

$$\min \sum_{i=1}^k \sum_{j=1}^h (y_j - y_j^*)^2 n_{ij} = \min \sum_{i=1}^k \sum_{j=1}^h (y_j - a - bx_i)^2 n_{ij}$$

$$b = \frac{s_{XY}}{s_X^2}$$

$$a = \bar{y} - b\bar{x}$$

la ecuación de la recta buscada puede ser expresada como sigue:

$$y - \bar{y} = \frac{s_{XY}}{s_X^2} (x - \bar{x})$$

(b) Regresión de X sobre Y

$$x_i^* = a' + b'y_j$$

$$\min \sum_{i=1}^k \sum_{j=1}^h (x_i - x_i^*)^2 n_{ij} = \min \sum_{i=1}^k \sum_{j=1}^h (x_i - a' - b'y_j)^2 n_{ij}$$

$$b' = \frac{s_{XY}}{s_Y^2}$$

$$a' = \bar{x} - b'\bar{y}$$

la ecuación de la recta buscada puede escribirse:

$$x - \bar{x} = \frac{s_{XY}}{s_Y^2} (y - \bar{y})$$

8.3.2. Propiedades de las rectas de regresión

Definición 8.3 Se llaman coeficientes de regresión lineal a las pendientes de las rectas de regresión.

El coeficiente de regresión lineal de Y sobre X , $b = \frac{s_{XY}}{s_X^2}$, mide la variación de la variable Y para un incremento unitario de X .

Análogamente, el coeficiente de regresión lineal de X sobre Y , $b' = \frac{s_{XY}}{s_Y^2}$, expresa la variación de la variable X para un incremento unitario de Y .

OBSERVACIÓN 8.3 Como podemos observar de las expresiones de b y b' , se obtiene que:

$$\text{signo}(b) = \text{signo}(b') = \text{signo}(s_{XY}),$$

y como consecuencia

- (a) Si $s_{XY} > 0$ (dependencia lineal directa) $\Rightarrow b, b' > 0$, y por tanto, ambas rectas son crecientes.
- (b) Si $s_{XY} < 0$ (dependencia lineal inversa) $\Rightarrow b, b' < 0$, y entonces, ambas rectas son decrecientes.
- (c) Si $s_{XY} = 0$ (variables incorreladas) $\Rightarrow b = b' = 0$. Las rectas son perpendiculares.

Definición 8.4 El punto de corte de las dos rectas de regresión $(\bar{x}, \bar{y})$ se llama centro de gravedad de la distribución.

EJEMPLO 8.3 Para 23 tiendas se han observado el precio, X , y el volumen de ventas anual, Y , en millones de unidades, de cierto artículo:

x_i	y_j	n_{ij}
10	12	3
11	12	4
12	10	5
13	10	7
14	8	3
15	6	1

(a) Calcula las rectas de regresión que relacionan las variables precio y volumen de ventas.

x_i	y_j	6	8	10	12	$n_{i.}$	$x_i n_{i.}$	$x_i^2 n_{i.}$
10		0	0	0	3	3	30	300
11		0	0	0	4	4	44	484
12		0	0	5	0	5	60	720
13		0	0	7	0	7	91	1183
14		0	3	0	0	3	42	588
15		1	0	0	0	1	15	225
$n_{.j}$		1	3	12	7	$N = 23$	282	3500
$y_j n_{.j}$		6	24	120	84	234		
$y_j^2 n_{.j}$		36	192	1200	1008	2436		

$$r_{Y/X} \equiv y - \bar{y} = \frac{s_{XY}}{s_X^2}(x - \bar{x})$$

$$\bar{x} = \frac{\sum_{i=1}^k x_i n_{i.}}{N} = \frac{282}{23}$$

$$\bar{y} = \frac{\sum_{j=1}^h y_j n_{.j}}{N} = \frac{234}{23}$$

$$s_{XY} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^h x_i y_j n_{ij} - \bar{x} \cdot \bar{y} = \frac{2824}{23} - \frac{282}{23} \cdot \frac{234}{23} = -\frac{1036}{529}$$

$$s_X^2 = \frac{1}{N} \sum_{i=1}^k x_i^2 n_{i.} - \bar{x}^2 = \frac{3500}{23} - \left(\frac{282}{23}\right)^2 = \frac{976}{529}$$

$$y = \frac{2829}{122} - \frac{259}{244}x$$

$$r_{X/Y} \equiv x - \bar{x} = \frac{s_{XY}}{s_Y^2} (y - \bar{y})$$

$$s_Y^2 = \frac{1}{N} \sum_{j=1}^h y_j^2 n_{.j} - \bar{y}^2 = \frac{2436}{23} - \left(\frac{234}{23} \right)^2 = \frac{1272}{529}$$

$$x - \frac{282}{23} = -\frac{259}{318} \left(y - \frac{234}{23} \right) \Rightarrow$$

$$x = \frac{1089}{53} - \frac{259}{318} y$$

- (b) *Calcula el número de unidades que se espera vender en una tienda si el precio sube a 17 unidades monetarias.*

$$y = \frac{2829}{122} - \frac{259}{244}(17) = \frac{1255}{244} = 5,143442 \text{ millones de artículos.}$$

- (c) *Calcula el precio estimado para conseguir unas ventas de 15 millones de unidades.*

$$x = \frac{1089}{53} - \frac{259}{318}(15) = \frac{2649}{318} = 8,330188 \text{ unidades monetarias.}$$

- (d) *¿En cuánto se estima que se reduzca el volumen de ventas ante un incremento unitario en el precio de dicho artículo?*

$$b = \frac{-259}{244} = -1,061475 \text{ millones de unidades.}$$

8.4. Línea de Tukey o línea Mediana-Mediana

Cuando se presume la existencia de valores anómalos (outliers) entre los datos que se manejan, no conviene utilizar la regresión mínimo-cuadrática, ya que podría quedar gravemente afectada por los citados valores. No olvidemos que la regresión está basada en las medias de las distribuciones condicionadas. En estos casos se utilizará la línea de Tukey que se basa en la mediana, más robusta e inerte a los datos anómalos que la media.

Para su cálculo seguiremos los siguientes pasos:

- (a) Ordenamos los pares de menor a mayor en base a la variable independiente.
- (b) Dividimos los datos en tres partes de modo que cada uno de ellos abarque aproximadamente la tercera parte de los datos.

- (c) Para el primer grupo de valores x^1 e y^1 calculamos sus medianas $M_e(x^1)$ y $M_e(y^1)$.
- (d) Análogamente se procede con el tercer grupo de datos, calculándose $M_e(x^3)$ y $M_e(y^3)$.
- (e) La línea de Tukey se obtiene trazando la recta que tiene la dirección dada por los puntos $(M_e(x^1), M_e(y^1))$ y $(M_e(x^3), M_e(y^3))$, y ordenada en el origen $M_e(y_i - bx_i)$.

Analíticamente la línea de Tukey será $y = a + bx$

$$b = \frac{M_e(y^3) - M_e(y^1)}{M_e(x^3) - M_e(x^1)}$$

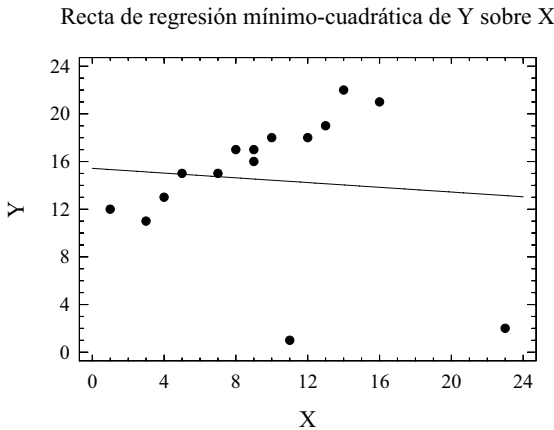
$$a = M_e(y_i - bx_i)$$

EJEMPLO 8.4

x_i	1	3	4	5	7	8	9	9	10	11	12	13	14	16	23
y_j	12	11	13	15	15	17	16	17	18	1	18	19	22	21	2

(a) Recta de regresión mínimo-cuadrática de Y sobre X

$$y = \frac{101667}{6590} - \frac{131}{1318}x = 15,427465 - 0,099393x$$



(b) Línea de Tukey

$$x^1 : 1, 3, 4, 5, 7 \Rightarrow M_e(x^1) = 4$$

$x^3 : 12, 13, 14, 16, 23 \Rightarrow M_e(x^3) = 14$

$y^1 : 12, 11, 13, 15, 15 \Rightarrow M_e(y^1) = 13$

$y^3 : 18, 19, 22, 21, 2 \Rightarrow M_e(y^3) = 19$

$$b = \frac{19 - 13}{14 - 4} = 0,6$$

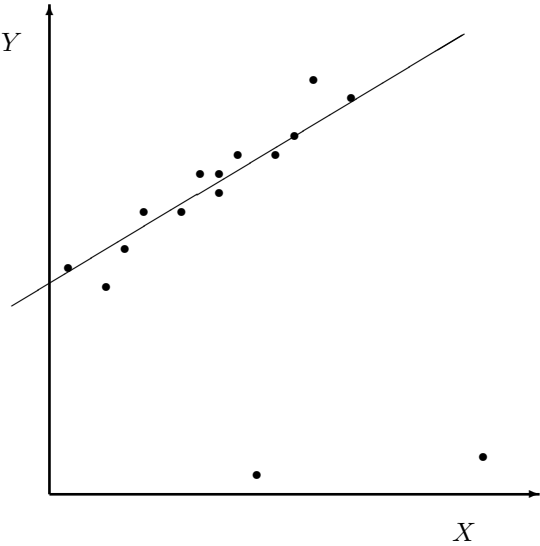
$y_i - 0,6x_i$		
11,4	12,2	10,8
9,2	10,6	11,2
10,6	11,6	13,6
12	12	11,4
10,8	-5,6	-11,8

Ordenados:

-11,8; -5,6; 9,2; 10,6; 10,6; 10,8; 10,8; 11,2; 11,4; 11,4; 11,6; 12; 12; 12,2; 13,6

$$a = M_e(y_i - 0,6x_i) = 11,2$$

La línea de Tukey es $y = 11,2 + 0,6x$



Capítulo 9

Correlación simple

9.1. Concepto de correlación

Definición 9.1 *El objetivo de la correlación es estudiar el grado de asociación existente entre las variables, es decir, proporcionar unos coeficientes que nos midan el grado de dependencia mutua entre las variables.*

Diremos que la dependencia es "*perfecta*" o que existe una dependencia "*funcional*" entre las variables si todos los puntos del diagrama de dispersión se encuentran sobre la línea de regresión. Cuánto más lejos se encuentren dichos puntos de la línea de regresión, menor será la intensidad de la dependencia entre las variables consideradas.

Toda línea de regresión debe ir acompañada de una medida de la "*bondad del ajuste*" o "*representatividad*", al igual que ocurría en el caso unidimensional, donde las medidas de posición se acompañaban de medidas de dispersión al objeto de ver la representatividad de las primeras.

9.2. Medidas de correlación

9.2.1. Varianza residual. Análisis de los residuos

La bondad o aproximación de un ajuste o regresión a los datos reales no pueden medirse por la suma de residuos porque errores positivos se podrían compensar con errores negativos, y además, en algunos procedimientos, como el de mínimo cuadrados que estamos considerando, dicha expresión es cero. Por lo tanto, tomaremos la suma de residuos al cuadrado, también llamada *Error Cuadrático Medio*, *E.C.M.*, como medida de bondad de ajuste. En funciones lineales en los parámetros, la media de los residuos es cero, por lo que el *E.C.M.* coincide con la varianza de los residuos o varianza residual.

Definición 9.2 La varianza residual es el coeficiente que mide la variabilidad de los residuos o errores y viene dada por la expresión

$$E.C.M = s_{rY}^2 = \sum_{i=1}^k \sum_{j=1}^h (y_j - y_j^*)^2 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^k \sum_{j=1}^h e_j^2 \cdot \frac{n_{ij}}{N}$$

siendo los los valores $e_j = y_j - y_j^*$, es decir, los residuos o errores.

- (a) Valores grandes de s_{rY}^2 indican, que en promedio, los valores $e_j = y_j - y_j^*$ son grandes, y como consecuencia, la línea de regresión es poco representativa.
- (b) Valores pequeños de s_{rY}^2 indicarían, que en promedio, los valores e_j son pequeños, y por tanto, la línea de regresión es representativa. La máxima representatividad se tiene si $e_j = 0$ para todo j , es decir, cuando $s_{rY}^2 = 0$, que es el mínimo valor que la varianza residual puede alcanzar.

Inconvenientes de la varianza residual

- (a) La varianza residual tiene el inconveniente de que depende de las unidades de medida al cuadrado. Esto hace que no sea posible comparar el grado de dependencia entre grupos de variables expresadas en distintas unidades de medida. Necesitamos por tanto una medida adimensional.
- (b) No nos dice a partir de qué valores es suficientemente pequeña o grande como para admitir un buen o mal ajuste.

9.2.2. Coeficiente de determinación. Interpretación

Definición 9.3 Se define el coeficiente de determinación como

$$R^2 = 1 - \frac{s_{rY}^2}{s_Y^2}$$

Ventajas del coeficientes de determinación

- (a) Al estar R^2 definido por cociente entre varianzas es, por tanto, un parámetro independiente de las unidades de medida. Esto nos permitirá efectuar comparaciones.
- (b) Su rango de variación es acotado, $0 \leq R^2 \leq 1$, ya que se verifica que $0 \leq s_{rY}^2 \leq s_Y^2$.
 - (i) Si el ajuste es perfecto, es decir, todos los puntos del diagrama de dispersión se sitúan sobre la curva calculada ($s_{rY}^2 = 0$), se verifica que $R^2 = 1$.

- (ii) Cuanto mayor sea la distancia de los puntos a la curva, mayor es $s_{r_Y}^2$ y menor R^2 . El valor mínimo de éste, $R^2 = 0$, se alcanza cuando $s_{r_Y}^2 = s_Y^2$, en cuyo caso no se consigue ninguna explicación de la variable Y relacionándola con la X mediante la curva considerada.

OBSERVACIÓN 9.1 *Cuando el coeficiente de determinación vale como mínimo 0,75, el modelo ajustado suele aceptarse. Si el coeficiente es inferior a dicho valor, concluiremos que la función elegida no es adecuada, debiéndose ensayar con otro tipo de función.*

9.3. Correlación lineal

9.3.1. Descomposición de la varianza total

Puede demostrarse que, en el caso lineal, es posible descomponer la varianza total en la forma siguiente:

$$s_Y^2 = s_R^2 + s_{r_Y}^2$$

siendo

$$s_Y^2 = \text{varianza total} = \sum_{i=1}^k \sum_{j=1}^h (y_j - \bar{y})^2 \cdot \frac{n_{ij}}{N}$$

$$s_R^2 = \text{varianza debida a la regresión} = \sum_{i=1}^k \sum_{j=1}^h (y_j^* - \bar{y}^*)^2 \cdot \frac{n_{ij}}{N}$$

La varianza debida a la regresión nos indica en qué medida queda explicada la variable dependiente mediante el modelo estimado, midiendo la dispersión de los valores en la recta. Y $s_{r_Y}^2$ es la conocida varianza residual que, en el caso lineal, al ser $\bar{e} = 0$, adopta la expresión de una varianza, llamada varianza debida al error y que puede escribirse como sigue:

$$s_{r_Y}^2 = \sum_{i=1}^k \sum_{j=1}^h (e_j - \bar{e})^2 \cdot \frac{n_{ij}}{N}$$

La varianza residual se puede interpretar como una medida de lo que queda sin explicar después de haber efectuado la regresión. La varianza residual debe recoger las variables no incluidas en el modelo y las deficiencias de especificación (por ejemplo, formular un modelo lineal cuando no debería haber sido exponencial). Mide, por tanto, las desviaciones entre los valores teóricos y los observados.

Por tanto, se verifica:

$$R^2 = 1 - \frac{s_{r_Y}^2}{s_Y^2} = \frac{s_Y^2 - s_{r_Y}^2}{s_Y^2} = \frac{s_R^2}{s_Y^2},$$

que expresa el porcentaje de la variabilidad de la variable Y que queda explicado por la regresión.

9.3.2. El coeficiente de correlación lineal

La covarianza es un estadístico cuyo valor depende de las unidades en que vengan expresadas las variables y se pretende buscar un indicador de esta relación o covariación al que no le afecte este problema.

Definición 9.4 *Se define el coeficiente de correlación lineal como*

$$r = \frac{s_{XY}}{s_X s_Y}$$

Este coeficiente nos da el grado de asociación lineal entre las variables, y el tipo de dicha relación.

Relación entre r y R^2

$$\begin{aligned} s_{r_Y}^2 &= \sum_{i=1}^k \sum_{j=1}^h (y_j - y_j^*)^2 \frac{n_{ij}}{N} = \sum_{i=1}^k \sum_{j=1}^h \left[y_j - \left(\bar{y} + \frac{s_{XY}}{s_X^2} (x_i - \bar{x}) \right) \right]^2 \frac{n_{ij}}{N} = \\ &= \sum_{i=1}^k \sum_{j=1}^h \left((y_j - \bar{y}) - \frac{s_{XY}}{s_X^2} (x_i - \bar{x}) \right)^2 \frac{n_{ij}}{N} = \\ &= \sum_{i=1}^k \sum_{j=1}^h (y_j - \bar{y})^2 \frac{n_{ij}}{N} - 2 \cdot \frac{s_{XY}}{s_X^2} \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})(y_j - \bar{y}) \frac{n_{ij}}{N} + \\ &+ \left(\frac{s_{XY}}{s_X^2} \right)^2 \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})^2 \frac{n_{ij}}{N} = s_Y^2 - 2 \cdot \frac{s_{XY}^2}{s_X^2} + \left(\frac{s_{XY}}{s_X^2} \right)^2 s_X^2 = s_Y^2 - \frac{s_{XY}^2}{s_X^2} \end{aligned}$$

luego

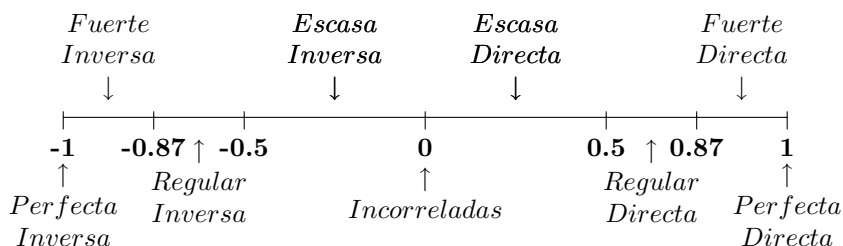
$$R^2 = 1 - \frac{s_{r_Y}^2}{s_Y^2} = 1 - \frac{s_Y^2 - \frac{s_{XY}^2}{s_X^2}}{s_Y^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2} = r^2$$

OBSERVACIÓN 9.2

- (a) Luego, en el caso lineal, se verifica que $R^2 = r^2$.
- (b) Aunque el coeficiente de determinación y el de correlación están relacionados, la aplicación que se da a cada uno es diferente. El primero se utiliza para medir la bondad de ajuste efectuado mediante la recta y el segundo para medir el tipo y grado de relación lineal entre las dos variables.

Interpretación de r

Como se indicó anteriormente, el coeficiente de correlación lineal r indica el tipo y grado o fortaleza de la dependencia lineal. Al verificarse que $R^2 = r^2$ y que $0 \leq R^2 \leq 1$, entonces $-1 \leq r \leq 1$. El signo hace alusión al tipo (directa o inversa) y su valor en términos absolutos, a la intensidad de la relación.



Propiedades de r

- r mide la apertura existente entre las rectas de regresión, siendo ambas rectas perpendiculares si $r = 0$, y coincidentes si $r = \pm 1$.
- Si X e Y son incorreladas, es decir, no existe dependencia lineal, se verifica que $r = 0$.
- Al ser $b = \frac{s_{XY}}{s_X^2}$ y $b' = \frac{s_{XY}}{s_Y^2}$, se tiene $r = \sqrt{b \cdot b'}$.
- $\text{signo}(b) = \text{signo}(b') = \text{signo}(s_{XY}) = \text{signo}(r)$.
- El coeficiente de correlación lineal es una medida resumen de la estructura de un diagrama de dispersión y que, en consecuencia, siempre conviene dibujar el diagrama de dispersión ya que éste sí contiene toda la información. Coeficientes de correlación cercanos a cero se pueden corresponder con diagramas de dispersión distintos.

9.4. Bondad de ajuste para otras funciones

Para otro tipo de funciones como la exponencial, potencial o hiperbólica, la varianza residual necesaria para el cálculo del coeficiente de determinación, se calculará utilizando la definición inicial de la misma, es decir

$$s_{rY}^2 = \sum_{i=1}^k \sum_{j=1}^h (y_j - y_j^*)^2 \frac{n_{ij}}{N}$$

donde $y_j^* = ab^{x_i}$, $y_j^* = ax_i^b$ ó $y_j^* = a + b \cdot \frac{1}{x_i}$, respectivamente.

9.5. Predicción

Uno de los objetivos que persigue la regresión es conseguir realizar predicciones sobre la variable dependiente basadas en los valores que toma la variable independiente.

La predicción será más fiable cuanto mayor sea el correspondiente coeficiente de determinación. Debemos también observar que los valores de la variable independiente para los que deseamos obtener predicciones no se encuentren muy alejados de su rango de variación, ya que en ese caso la predicción pudiera no ser fiable.

9.6. Correlación espuria

Esta correlación se produce al seleccionar dos variables cualesquiera al azar y que por casualidad el coeficiente de correlación en valor absoluto sea uno o próximo a uno, pero sin que exista verdaderamente ningún tipo de relación entre ellas, como por ejemplo: X ="consumo de gasolina" e Y ="consumo de bebidas", dependen ambas de una tercera, la renta.

Otro ejemplo puede ser la existencia de una correlación positiva y fuerte entre las variables X ="ventas de chicle" e Y ="incidencia del crimen en los Estados Unidos". Obviamente no es lícito que prohibiendo el consumo de chicle se reduzca la tasa de criminalidad. Ambas variables dependen de una tercera: el tamaño de la población analizada.

Cuestiones, problemas y recursos

Cuestiones

1. Dada la distribución bidimensional recogida en la siguiente tabla,

$X \backslash Y$	1	2	3
10	1	4	5
11	0	2	3
12	3	1	1

calcula la distribución condicionada $X/Y = 2$ y la de $Y/X = 12$.

2. La distribución de la edad, X , y la antigüedad, Y , de los trabajadores de un departamento comercial es la siguiente:

$X \backslash Y$	7-10	10-13	13-16	16-19
49-52	0	2	8	6
52-55	3	8	9	2
55-58	7	5	0	0

De los trabajadores que llevan más de 13 años en dicho departamento comercial, calcula qué porcentaje tiene una edad superior a 54 años.

3. Dada la distribución bidimensional

$X \backslash Y$	2	0	2	0
Y	0	2	0	-2

¿Pueden considerarse incorreladas las variables X e Y ? ¿Pueden considerarse independientes?

4. Representa adecuadamente el diagrama de dispersión de una distribución bidimensional donde las variables que la componen sean dependientes e incorreladas. Razona la respuesta.
5. En un ajuste lineal $r = -0,7$. ¿Qué porcentaje de la variabilidad de la variable dependiente queda explicada por la regresión lineal?
6. ¿Qué puntos constituyen la regresión de la media de Y sobre X en la siguiente distribución:?

$X \backslash Y$	1	2	3
1	2	2	1
2	3	2	1
3	4	2	1

7. En una distribución bidimensional las rectas de regresión son $x = 4$, $y = 2$. ¿Se puede decir que no existe ningún tipo de relación entre las variables? Razona la respuesta.
8. En una distribución bidimensional, ¿son compatibles los siguientes datos $s_{XY} = -4$, $y = 3 + 2x$, $r = -0,9$? Razona la respuesta.
9. De una distribución bidimensional se sabe que la regresión lineal realizada consigue explicar el 64 % de la variabilidad de la variable dependiente Y . Sabiendo que $s_Y^2 = 64$ y $r_{Y/X} : y = 3 - 1,6x$, calcula s_{XY} .
10. Dada una distribución bidimensional se ha calculado la recta de regresión mínimo cuadrática de Y sobre X , obteniéndose: $r_{Y/X} \equiv y = 2x + 1$. Se sabe, además, que dicha regresión consigue explicar el 96 % de la variabilidad de la variable Y .
 - (a) Calcula $r_{X/Y}$ sabiendo que el centro de gravedad de la distribución es $(\bar{x}, \bar{y}) = (1, 3)$.
 - (b) Tipo y grado de la relación lineal entre las variables.
11. Se desea calcular la ecuación de una recta que explique el comportamiento de la variable estadística Y en función de la variable X , para la siguiente distribución bidimensional:

X	5	5	6	7	90
Y	3.5	3	2.5	2	1

¿Qué procedimientos conoces para la obtención de dicha recta? Razona cuál es el más adecuado para esta distribución concreta.

12. Dada una distribución bidimensional formada por 10 pares de valores se han calculado los datos: $\bar{x} = 2$, $\bar{y} = 5$, $\sum x_i^2 = 80$, $\sum y_j^2 = 280$ y $\sum \sum x_i y_j = 70$. Calcula $r_{Y/X}$. ¿Es directa la dependencia entre las variables?
13. El coeficiente de correlación de una distribución bidimensional es $r = -0,98$. Interpreta adecuadamente dicho valor y dibuja aproximadamente la nube de puntos y las rectas de regresión que se correspondan con el valor anterior.
14. ¿Qué es el centro de gravedad de una distribución bidimensional? Calcula el centro de gravedad de la distribución a partir de la cual se obtuvieron las siguientes rectas de regresión $y = 2x + 1$ y $x = 4y - 11$.
15. En una distribución bidimensional, al realizar la recta de regresión $y = a + bx$, se han obtenido los siguientes datos: $s_{XY} = 20$, $s_X^2 = 10$, $a = 3$, $\bar{y} = 8$, $\bar{x} = 4$. ¿Son compatibles? Razona la respuesta.
16. Si la recta de regresión de Y sobre X es decreciente y coincide con la recta de regresión de X sobre Y , razona la veracidad o falsedad de cada una de las siguientes afirmaciones:
 - (a) El coeficiente de correlación lineal entre X e Y vale -1.
 - (b) El coeficiente de determinación del modelo lineal vale -1.
 - (c) El coeficiente de determinación del modelo lineal vale 0.
17. A partir de cierta distribución bidimensional se han calculado las rectas de regresión mínimo-cuadráticas obteniéndose: $x = -4y + 1$ e $y = -\frac{1}{4}x + \frac{1}{4}$. Realiza un comentario detallado acerca de la posible relación lineal existente entre ambas variables.

Problemas

1. Estudiada una muestra de cien contribuyentes de cierta ciudad se ha obtenido la siguiente información sobre la renta, (X), y la cantidad pagada en concepto de impuestos, (Y), estando ambas variables expresadas en cientos de miles de unidades monetarias. Los resultados observados se reflejan en la siguiente tabla:

Y			
X	0 – 2	2 – 4	4 – 6
6 – 10	20	0	0
10 – 20	5	30	5
20 – 40	0	20	20

- Calcula la covarianza. ¿Pueden considerarse independientes ambas variables?
 - ¿Qué valor puede considerarse más representativo, la renta media o la cantidad media pagada en impuestos?
 - ¿Cuál es el valor de la renta más frecuente?
 - Si tenemos en cuenta los contribuyentes con renta comprendida entre un millón y dos millones de u.m., ¿qué cantidad pagan como mínimo en impuestos el 40 % de los que más pagan?
 - Realiza una representación gráfica adecuada de la variable renta para aquellos contribuyentes que pagan en impuestos entre 200000 y 400000 unidades monetarias.
2. Dada la siguiente distribución bidimensional:

X	2	3	4	5
Y	3	5	4	6

Se pide:

- La recta de regresión de Y sobre X .
 - La recta de regresión de X sobre Y .
 - El tipo y la fortaleza de la relación de las variables.
 - ¿Qué valor es el predicho para la Y si $X = 30$? Comenta los argumentos a favor y en contra sobre la fiabilidad de dicho resultado.
3. Dada la siguiente distribución bidimensional:

X	1	2	4	5	6
Y	1	2	2	4	50

Se pide:

- (a) La recta de regresión de Y sobre X .
- (b) El tipo y grado de relación de las variables.
- (c) ¿Qué tanto por ciento de la varianza total es explicada por la regresión?
- (d) La línea de Tukey. Interpretación.

4. Dada la siguiente distribución bidimensional:

X	1	2	3	4	5
Y	3	3	5	6	5
n_{ij}	2	4	3	1	3

Se pide:

- (a) La recta de regresión de Y sobre X .
- (b) Valor pronosticado para la X si $Y = 7$.
- (c) La varianza residual. Interpretación.
- (d) El coeficiente de determinación. Interpretación.

5. Las notas obtenidas por 10 universitarios en dos asignaturas han sido:

X=Estadística	6	8	7	5	0
Y=Derecho	7	8	9	6	1
n_{ij}	2	4	2	1	1

- (a) A través de un ajuste lineal, se pide la nota pronosticada en Derecho para un alumno que en Estadística sacó un 9.
 - (b) Haz lo mismo que en el anterior apartado pero suprimiendo el último alumno que no estudió.
 - (c) Compara los coeficientes de determinación de los apartados anteriores.
 - (d) ¿Puedo suponer con razón que a mayor nota en Estadística mayor nota en Derecho?
6. En cierta empresa, para analizar la eficacia de la publicidad sobre las ventas de un determinado artículo, se han anotado los gastos en miles de euros, (X), a lo largo de cinco campañas publicitarias y el volumen de ventas de dicho artículo en miles de euros, (Y), en el periodo posterior a cada campaña, obteniéndose la siguiente tabla:

X	12	18	30	36	60
Y	300	360	720	900	1080

Se pide:

- (a) La recta de regresión que explique las ventas en función de los gastos en publicidad.
 - (b) El porcentaje de la variabilidad del volumen de ventas explicado por los gastos en publicidad.
 - (c) El volumen de ventas que se espera para un gasto en publicidad de 50000 euros. Razona si puede considerarse fiable tal predicción.
7. De una muestra de 14 puestos de venta del mercado de abastos de Jerez se recogían información del número de balanzas, (X), y número de dependientes, (Y),

Y X	1	2	3
1	1	2	0
2	1	2	3
3	0	1	2
4	0	0	2

- (a) Calcula la recta que explica el número de balanzas en función del número de dependientes.
 - (b) Calcula el tipo de relación entre esas dos variables a través de dos coeficientes distintos.
 - (c) ¿Cuál es el número de dependientes que tendrá una tienda si el número de balanzas es de 7? ¿Es fiable tal predicción?
8. Dada la siguiente distribución bidimensional, se pide:

Y=Demanda X=Precio	40-50	50-80	80-90
1	2	3	0
2	0	2	1
7	0	0	1
9	0	0	1

- (a) La recta de regresión de Y sobre X y la de X sobre Y .
 - (b) El coeficiente de correlación y el de determinación. Interpretación.
 - (c) ¿Cuál sería la demanda esperada si el precio es de 10 pesetas?
9. Realizado un test de conducción a 40 personas para ver la relación entre la edad y los reflejos se obtuvieron los siguientes resultados:

Puntuación Edad	60-90	90-120	120-150
10-20	3	2	0
20-30	4	6	6
30-40	11	4	3
40-50	1	0	0

- (a) Calcula las dos rectas de regresión.
- (b) ¿Cuál sería la puntuación obtenida por un jubilado de 60 años?
- (c) ¿Cuál sería el tipo y grado de relación lineal que existe entre las variables en estudio?
- (d) ¿Qué edad tendría una persona que obtuvo una puntuación de 160? ¿Sería fiable tal predicción?
10. Los inversores en acciones tienden a diversificar sus carteras de valores. Una entidad bancaria desea comprobar si esa tendencia se da entre ocho de sus clientes, y para ello recaba información sobre los diferentes tipos de acciones que tienen, (X), y los millones de unidades monetarias que han invertido en las mismas, (Y):

X	12	8	10	11	7	7	10	14
Y	58	42	51	54	40	39	49	56

- (a) Comenta razonadamente si existe asociación lineal entre ambas variables.
- (b) Estudia la bondad del modelo lineal para explicar el comportamiento de una de las variables en función de la otra.
- (c) Si el modelo lineal fuese adecuado para hacer predicciones, ¿cuántos millones de u.m. podría invertir un cliente que quisiera diversificar su cartera con nueve tipos de acciones?
- (d) Realiza un ajuste hiperbólico que exprese el comportamiento de la variable Y en función de X .
- (e) ¿Cómo ha variado la bondad del ajuste con el nuevo modelo?
11. Se dispone de los siguientes datos sobre los costes de producción por unidad, (Y), y los diferentes niveles de producción correspondientes a los mismos, (X), de cierta aserradora, ambas medidas en miles de unidades:

Nivel de producción	1	1.5	1.8	2.1	3
Coste unitario	3.25	1.25	1	0.825	0.45

- (a) Expresa el coste unitario en función del nivel de producción a través de un ajuste hiperbólico.

- (b) Estudia la bondad del ajuste realizado. Interpretación.
- (c) Estima el coste unitario para un nivel de producción de 2500 unidades.
12. Ajustar a una función de Makeham los siguientes datos del número de individuos existentes en una población en ciertos años por el método de las sumas:

X	1995	1996	1997	1998	1999	2000
Y	100	90	100	110	130	110

13. Ajustar a una logística los siguientes datos que reflejan las ventas en millones de euros de una empresa por el método de los valores recíprocos:

X	1999	2000	2001	2002	2003	2004
Y	40.3	47.6	55.1	62.4	69.1	75.1

Recursos

Seguidamente exponemos una serie de páginas de la web donde podemos encontrar algunos recursos electrónicos relacionados con esta parte:

- (a) [http : //www.coe.tamu.edu/ ~ strader/math166H/LeastSquares/ls2.html](http://www.coe.tamu.edu/~strader/math166H/LeastSquares/ls2.html) Permite trabajar con correlación utilizando tablas o directamente desde la nube de puntos.
- (b) [http : //www.stat.sc.edu/ ~ west/javahtml/Regression.html](http://www.stat.sc.edu/~west/javahtml/Regression.html) de la Universidad de Carolina del Sur, muestra cómo va cambiando la recta ajustada al añadir un nuevo punto.
- (c) [http : //www.ruf.rice.edu/ ~ lane/stat_sim/reg_by_eye/index.html](http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html) página de la Universidad de Rice (Tejas), presenta una nube de puntos a la que hay que ajustar una recta a ojo. También se puede intentar acercar el valor del coeficiente de correlación lineal, entre cinco propuestos. Finalmente, se puede comparar el ajuste propuesto con el obtenido por mínimos cuadrados.
- (d) [http : //www.stat.uiuc.edu/ ~ stat100/java/guess/GCApplet.html](http://www.stat.uiuc.edu/~stat100/java/guess/GCApplet.html) genera cuatro diagramas con coeficientes de correlación aleatorios y plantea el juego de intentar adivinar cuál es el coeficiente de correlación que corresponde a cada uno. También se generan diagramas con coeficientes de correlación impuestos.
- (e) [http : //www.duxbury.com/authors/mcclellandg/tiein/johnson/reg.html](http://www.duxbury.com/authors/mcclellandg/tiein/johnson/reg.html) al girar la recta muestra como va cambiando la suma de residuos y el valor del coeficiente de determinación.

Parte III

NÚMEROS ÍNDICES Y SERIES TEMPORALES

Capítulo 10

Números índices

10.1. Concepto de número índice

EJEMPLO 10.1 *La siguiente tabla presenta los precios del litro de una determinada marca de leche, en pesetas, en el período 1993-1998:*

Años	Precios/litro
1993	78
1994	83
1995	90
1996	92
1997	95
1998	99

Pretendemos estudiar la evolución en el tiempo del precio del litro de leche. Para ello podemos comparar el precio de cada año con el de uno fijo que se considera como referencia. Esta comparación puede realizarse de dos formas distintas:

(a) *Por diferencia:*

$$D = \text{Precio año } t - \text{Precio año de referencia}$$

En el ejemplo 10.1, el incremento del precio del litro de leche correspondiente al año 1998, respecto del 1993 ha sido

$$99 \text{ pesetas/litro} - 78 \text{ pesetas/litro} = 21 \text{ Pesetas/litro}$$

(b) *Por cociente:*

$$C = \frac{\text{Precio año } t}{\text{Precio año de referencia}}$$

En este caso, el litro de leche ha incrementado su precio en un 26,92 %, ya que

$$C = \frac{99 \text{ pesetas/litro}}{78 \text{ pesetas/litro}} = 1,2692 \Rightarrow 126,92 \%.$$

OBSERVACIÓN 10.1 *Debido a los problemas fundamentalmente derivados de las unidades de medida que presentaría la comparación por diferencia, efectuaremos siempre comparaciones por cociente. Esto nos permitirá agregar distintos índices para construir indicadores de evolución general de fenómenos económicos.*

Definición 10.1 *Se llama número índice a aquella medida estadística que nos permite estudiar la variación relativa que se produce en una magnitud a lo largo del tiempo o del espacio. Lo más habitual es estudiar la evolución de la magnitud a lo largo del tiempo, y ello se realiza mediante la comparación del valor de la magnitud en dos períodos temporales distintos.*

Definición 10.2 *Al período respecto del cual realizaremos las comparaciones se le llama período base o de referencia.*

OBSERVACIÓN 10.2 *El valor de la magnitud respecto del cual se efectúan las comparaciones no debe estar afectado por ningún acontecimiento extraordinario acaecido en ese período, ya que el resto de los valores de la serie se expresarán como porcentaje de variación respecto a él. Si este valor de comparación fuese atípico, la información que aportan los números índices puede conducir a equívocos.*

Definición 10.3 *Al período que queremos comparar se le llama período actual o corriente.*

10.2. Índices simples

Definición 10.4 *Consideremos una magnitud X que posee una sola modalidad y de la que conocemos su valor, x_0 , en el período base, y su valor en el período t , x_t . Se define el índice simple de X como*

$$I_0^t = I_{0,t} = \frac{x_t}{x_0}$$

y nos mide la variación, expresada en tantos por uno, que ha sufrido la magnitud considerada entre los períodos base y actual.

En el ejemplo 10.1,

$$I_{93}^{98} = I_{93,98} = \frac{99}{78} = 1,2692 \Rightarrow 126,92 \%$$

Los números índices suelen presentarse multiplicados por cien y expresados, por tanto, en porcentajes.

Los índices simples más usuales son aquellos que utilizan como magnitud el precio, la cantidad producida o vendida de un cierto bien, o el valor del mismo.

10.3. Propiedades de los índices simples

Los índices simples verifican las siguientes propiedades:

- (a) Existencia: El índice debe poder calcularse para cualquier valor de la magnitud.
- (b) Identidad: Si se hacen coincidir los períodos base y actual, el índice debe valer uno, o cien si se expresa de forma porcentual.

$$I_{0,0} = 1 \Rightarrow 100\%$$

- (c) Proporcionalidad: Si todos los valores de la magnitud son incrementados en un determinado porcentaje en el período actual, el índice debe variar en la misma proporción.

$$I'_{0,t} = \frac{x'_t}{x_0} = \frac{(1+k)x_t}{x_0} = (1+k) I_{0,t}$$

- (d) Transitividad o circularidad: El producto de dos índices para los que el período actual del primero coincida con el base del segundo, produce como resultado el índice que tiene como período base el del primero y como período actual el del segundo.

$$I_{0,t'} \cdot I_{t',t} = \frac{x_{t'}}{x_0} \cdot \frac{x_t}{x_{t'}} = I_{0,t}$$

- (e) Inversión: El producto de dos índices en los que se han invertido los períodos base y actual es igual a la unidad.

$$I_{0,t} \cdot I_{t,0} = \frac{x_t}{x_0} \cdot \frac{x_0}{x_t} = 1$$

10.4. Índices complejos

EJEMPLO 10.2 *La siguiente tabla presenta los precios de diferentes artículos en el período 1993-1998*

Años	litro/leche	kilo/pan	litro/cava
1993	78	72	425
1994	83	78	470
1995	90	82	510
1996	92	82	540
1997	95	83	545
1998	99	85	560

En la realidad sucede que, en muchos casos, estamos interesados en comparar precios, cantidades o valores para grupos de bienes.

Un planteamiento inicial podría consistir en sumar los precios de todos los productos para formar una serie única, y calcular los índices definidos anteriormente. Es decir, si quisiéramos estudiar la evolución entre los años 1993 y 1998 del precio del conjunto de bienes dado en el Ejemplo 10.2, obtendríamos

$$I_{93,98} = \frac{99 + 85 + 560}{78 + 72 + 425} = \frac{744}{575} = 1,2939 \Rightarrow 129,39\%$$

Sin embargo, un índice así construido se modifica cuando se cambian las unidades de medida.

Años	litro/leche	100 gramos de pan	litro/cava
1993	78	7,2	425
1994	83	7,8	470
1995	90	8,2	510
1996	92	8,2	540
1997	95	8,3	545
1998	99	8,5	560

$$I_{93,98} = \frac{99 + 8,5 + 560}{78 + 7,2 + 425} = \frac{667,5}{510,2} = 1,3083 \Rightarrow 130,83\%$$

Cuando se estudia la evolución de una magnitud que presenta más de una modalidad, como en el Ejemplo 10.2, una alternativa más razonable es construir en primer lugar los índices simples para cada uno de los productos, con igual período base, y promediarlos. Se obtiene así lo que se denomina índice complejo. Este procedimiento elimina el problema de las unidades en que se mide cada uno de los artículos.

Definición 10.5 *Un índice complejo es un promedio de un conjunto de índices simples y resume la variación global de una magnitud que presenta más de una modalidad.*

10.4.1. Índices complejos no ponderados

Consideremos la magnitud compleja X formada por las simples $X_1, X_2, \dots, X_N$ que han tomado los valores resumidos en el cuadro siguiente:

Período base	Período actual	Índices simples
x_{10}	x_{1t}	$I_{0,t}^1 = \frac{x_{1t}}{x_{10}}$
$\vdots$	$\vdots$	$\vdots$
x_{i0}	x_{it}	$I_{0,t}^i = \frac{x_{it}}{x_{i0}}$
$\vdots$	$\vdots$	$\vdots$
x_{N0}	x_{Nt}	$I_{0,t}^N = \frac{x_{Nt}}{x_{N0}}$

Según el promedio empleado obtenemos

(a) Índice aritmético complejo no ponderado

$$I_{A/0,t} = \frac{\sum_{i=1}^N I_{0,t}^i}{N}$$

(b) Índice geométrico complejo no ponderado

$$I_{G/0,t} = \sqrt[N]{\prod_{i=1}^N I_{0,t}^i}$$

(c) Índice armónico complejo no ponderado

$$I_{H/0,t} = \frac{N}{\sum_{i=1}^N \frac{1}{I_{0,t}^i}}$$

EJEMPLO 10.3 *Se desean calcular los índices complejos no ponderados anteriores para estudiar la evolución del precio del conjunto de bienes dado en el año 1998*

respecto del año 1993.

Años	litro/leche	kilo/pan	litro/cava
1993	78	72	425
1994	83	78	470
1995	90	82	510
1996	92	82	540
1997	95	83	545
1998	99	85	560

Los índices simples para estudiar la evolución individual del precio de cada artículo entre 1993 y 1998 son:

$I_{93,98}^i$	$\frac{99}{78} = 1,2692$	$\frac{85}{72} = 1,1805$	$\frac{560}{425} = 1,3176$
---------------	--------------------------	--------------------------	----------------------------

$$I_{A/93,98} = \frac{\sum_{i=1}^3 I_{93,98}^i}{3} = \frac{1,2692 + 1,1805 + 1,3176}{3} = 1,2557 \Rightarrow 125,57\%$$

$$I_{G/93,98} = \sqrt[3]{\prod_{i=1}^3 I_{93,98}^i} = \sqrt[3]{1,2692 \cdot 1,1805 \cdot 1,3176} = 1,2544 \Rightarrow 125,44\%$$

$$I_{H/93,98} = \frac{3}{\sum_{i=1}^3 \frac{1}{I_{93,98}^i}} = \frac{3}{\frac{1}{1,2692} + \frac{1}{1,1805} + \frac{1}{1,3176}} = 1,2531 \Rightarrow 125,31\%$$

10.4.2. Índices complejos ponderados

En todos los índices complejos anteriores no se ha tenido en cuenta la diferente importancia relativa que puede tener cada una de las magnitudes simples dentro del conjunto de todas ellas. Los índices complejos ponderados introducen *ponderaciones* que miden el peso relativo de cada magnitud simple dentro del conjunto en que se considere. Este tipo de índices son los que realmente se emplean en el estudio de la evolución de los fenómenos complejos de naturaleza económica: índice de precios de consumo (IPC), índice de producción industrial (IPI), índice de precios industriales (IPRI), etc.

(a) Índice aritmético complejo ponderado

$$I_{A/0,t}^* = \frac{\sum_{i=1}^N \omega_i \cdot I_{0,t}^i}{\sum_{i=1}^N \omega_i}$$

(b) Índice geométrico complejo ponderado

$$I_{G/0,t}^* = \left[\prod_{i=1}^N (I_{0,t}^i)^{\omega_i} \right]^{\frac{1}{\sum_{i=1}^N \omega_i}}$$

(c) Índice armónico complejo ponderado

$$I_{H/0,t}^* = \frac{\sum_{i=1}^N \omega_i}{\sum_{i=1}^N \frac{\omega_i}{I_{0,t}^i}}$$

EJEMPLO 10.4 *Se desean calcular los índices complejos ponderados anteriores para estudiar la evolución del precio del conjunto de bienes dado en el año 1998 respecto del año 1993. Para ello supondremos conocidas las ponderaciones de cada uno de los artículos.*

Años	1993	1998	ω_i
litro/leche	78	99	4
kilo/pan	72	87	5
litro/cava	425	560	1

Recordemos que los índices simples para estudiar la evolución individual del precio de cada artículo entre 1993 y 1998 son:

$I_{93,98}^i$	$\frac{99}{78} = 1,2692$	$\frac{87}{72} = 1,1805$	$\frac{560}{425} = 1,3176$
---------------	--------------------------	--------------------------	----------------------------

entonces,

$$I_{A/93,98}^* = \frac{\sum_{i=1}^3 \omega_i \cdot I_{93,98}^i}{\sum_{i=1}^3 \omega_i} = \frac{4 \times 1,2692 + 5 \times 1,1805 + 1 \times 1,3176}{4 + 5 + 1} = 1,2296$$

$$I_{G/93,98}^* = \left[\prod_{i=1}^3 (I_{93,98}^i)^{\omega_i} \right]^{\frac{1}{\sum_{i=1}^3 \omega_i}} = \sqrt[10]{1,2692^4 \cdot 1,1805^5 \cdot 1,3176^1} = 1,2286$$

$$I_{H/93,98}^* = \frac{\sum_{i=1}^3 \omega_i}{\sum_{i=1}^3 \frac{\omega_i}{I_{93,98}^i}} = \frac{10}{\frac{4}{1,2692} + \frac{5}{1,1805} + \frac{1}{1,3176}} = 1,2275$$

10.5. Índices encadenados

Se llama índice encadenado al que se obtiene a través de enlaces relativos. Es un índice resultado del producto de índices para los cuales la base es siempre el período precedente. Cada uno de ellos representa una comparación porcentual respecto al período anterior.

Definición 10.6 Si entre los períodos 0 y t pasamos por una serie de situaciones intermedias $1, 2, 3, \dots, (t-1)$, llamaremos índice encadenado según la fórmula k a

$$I_{k/0,t}^c = I_{k/0,1} \cdot I_{k/1,2} \cdot I_{k/2,3} \cdots I_{k/t-1,t}$$

Capítulo 11

Índices de precios

11.1. Introducción

Si la magnitud a analizar X es el precio, hablaremos de índices de precios.

Definición 11.1 Si X_i representa el precio del artículo o servicio i de la que conocemos su valor p_{i0} , en el período base, y su valor en el período t , p_{it} , se define el índice simple de precios para dicho artículo o servicio como

$$I_{0,t}^i = \frac{p_{it}}{p_{i0}}$$

Como se indicó en el capítulo precedente, los índices más comúnmente utilizados son los índices de precios complejos, que son promedios de los índices de precios simples. La idea es reflejar la evolución conjunta del precio de un grupo de artículos o servicios. Entre los índices de precios complejos, los más empleados son los ponderados ya que tienen en cuenta la importancia relativa que tiene cada artículo o servicio en el conjunto considerado. Los índices de precios complejos que estudiaremos aparecen resumidos en la siguiente tabla:

Índices complejos	
No ponderados	Ponderados
Sauerbeck-Carli	Laspeyres
Bradstreet-Dûtot	Paasche
	Marshall-Edgeworth-Bowley
	Fisher

11.2. Índices de precios complejos no ponderados

En este caso queremos estudiar la evolución conjunta del precio de un grupo de N componentes (artículos o servicios) otorgándole a cada uno de ellos la misma

importancia relativa. Los índices que veremos se definen empleando dos criterios: el de la media aritmética o el de la media agregativa.

(a) SAUERBECK-CARLI:

$$I_{A/0,t} = I_{S/0,t} = \frac{\sum_{i=1}^N I_{0,t}^i}{N} = \frac{1}{N} \sum_{i=1}^N \frac{p_{it}}{p_{i0}}$$

(b) BRADSTREET-DÛTOT:

$$I_{AG/0,t} = I_{BD/0,t} = \frac{\sum_{i=1}^N p_{it}}{\sum_{i=1}^N p_{i0}}$$

OBSERVACIÓN 11.1 *Este criterio de la media agregativa es aplicable exclusivamente en el caso de que trabajemos con magnitudes o variables expresadas en las mismas unidades de medida.*

EJEMPLO 11.1 *La tabla siguiente presenta los precios de cuatro artículos, medidos en las mismas unidades. Suponiendo que el año base sea 1996, calcula los dos índices complejos no ponderados anteriores para el año 1998 respecto del año 1996.*

Artículo	1996	1997	1998
A	30	32	35
B	80	84	89
C	200	220	235
D	900	1100	1250

Calculamos en primer lugar los índices de precios simples.

$I_{96,98}^i$	$\frac{35}{30} = 1,16$	$\frac{89}{80} = 1,11$	$\frac{235}{200} = 1,17$	$\frac{1250}{900} = 1,38$
---------------	------------------------	------------------------	--------------------------	---------------------------

$$I_{S/96,98} = \frac{\sum_{i=1}^4 I_{96,98}^i}{4} = \frac{1,16 + 1,11 + 1,17 + 1,38}{4} = 1,2107 \Rightarrow 121,07 \%$$

Este índice indica un incremento del precio del conjunto formado por esos cuatro artículos en un 21,07 %, en el año 1998 respecto de 1996.

$$I_{BD/96,98} = \frac{\sum_{i=1}^4 p_{i98}}{\sum_{i=1}^4 p_{i96}} = \frac{35 + 89 + 235 + 1250}{30 + 80 + 200 + 900} = 1,3297 \Rightarrow 132,97 \%$$

En este caso, el incremento del precio del conjunto formado por estos cuatro artículos es de un 32,97 % en el año 1998 respecto de 1996.

11.3. Índices de precios complejos ponderados

El inconveniente principal de los índices de precios complejos no ponderados es que no consideran la importancia relativa de cada elemento en el conjunto cuando, en los casos reales de los fenómenos económicos, esto debe ser tenido en cuenta. Es claro que, por ejemplo, en el presupuesto de un consumidor que normalmente gasta más en pan que en cava, ambos artículos no deben tener el mismo peso. Por tanto, parece lógico asignar diferentes pesos o ponderaciones, ω_i , a cada componente y, por extensión, a cada índice simple.

Tomando la definición de índice aritmético complejo ponderado, ya vista en el tema anterior, y utilizando como magnitud el precio, se tiene:

$$I_{A/0,t}^* = \frac{\sum_{i=1}^N \omega_i \cdot I_{0,t}^i}{\sum_{i=1}^N \omega_i} = \frac{\sum_{i=1}^N \omega_i \cdot \frac{p_{it}}{p_{i0}}}{\sum_{i=1}^N \omega_i} = \sum_{i=1}^N W_i \cdot \frac{p_{it}}{p_{i0}}$$

siendo

$$\omega_i = \text{ponderación o peso absoluto del componente } i.$$

$$W_i = \frac{\omega_i}{\sum_{i=1}^N \omega_i} \text{ ponderación o peso relativo del componente } i \left(\sum_{i=1}^N W_i = 1 \right).$$

En función de las ponderaciones consideradas, destacaremos los índices de precios siguientes:

- (a) LASPEYRES: utiliza como promedio la media aritmética ponderada, y como pesos las cantidades consumidas o vendidas de cada artículo o servicio

en el período base, valoradas a precios de dicho período ($\omega_i = p_{i0}q_{i0}$).

$$I_{L/0,t} = \frac{\sum_{i=1}^N \omega_i I_{0,t}^i}{\sum_{i=1}^N \omega_i} = \frac{\sum_{i=1}^N p_{i0}q_{i0} \frac{p_{it}}{p_{i0}}}{\sum_{i=1}^N p_{i0}q_{i0}} = \frac{\sum_{i=1}^N p_{it}q_{i0}}{\sum_{i=1}^N p_{i0}q_{i0}}$$

Este índice presenta la ventaja de que, fijado el período base, las ponderaciones se mantienen fijas para los diferentes períodos actuales considerados. El inconveniente que presenta es que pierde representatividad a medida que nos alejamos de dicho período base.

- (b) PAASCHE: utiliza como promedio la media aritmética ponderada, y como pesos las cantidades consumidas o vendidas de cada artículo o servicio en el período actual, valoradas a precios del período base ($\omega_i = p_{i0}q_{it}$).

$$I_{P/0,t} = \frac{\sum_{i=1}^N \omega_i I_{0,t}^i}{\sum_{i=1}^N \omega_i} = \frac{\sum_{i=1}^N p_{i0}q_{it} \frac{p_{it}}{p_{i0}}}{\sum_{i=1}^N p_{i0}q_{it}} = \frac{\sum_{i=1}^N p_{it}q_{it}}{\sum_{i=1}^N p_{i0}q_{it}}$$

Este índice presenta la ventaja de que las ponderaciones de los diferentes componentes se actualizan en cada período actual. El inconveniente que presenta es el aumento en complejidad y coste de su cálculo al tener que recoger información, no sólo sobre los precios, sino también sobre las cantidades de cada artículo o servicio en cada período considerado.

- (c) MARSHALL-EDGEWORTH-BOWLEY: en este caso también se utiliza como promedio la media aritmética ponderada. Los pesos considerados son las medias de las cantidades consumidas o vendidas en los períodos base y actual de cada artículo o servicio, valoradas a precios del período base ($\omega_i = p_{i0} \left(\frac{q_{i0} + q_{it}}{2} \right)$).

$$I_{MEB/0,t} = \frac{\sum_{i=1}^N \omega_i I_{0,t}^i}{\sum_{i=1}^N \omega_i} = \frac{\sum_{i=1}^N p_{i0} \left(\frac{q_{i0} + q_{it}}{2} \right) \frac{p_{it}}{p_{i0}}}{\sum_{i=1}^N p_{i0} \left(\frac{q_{i0} + q_{it}}{2} \right)} = \frac{\sum_{i=1}^N p_{it}(q_{i0} + q_{it})}{\sum_{i=1}^N p_{i0}(q_{i0} + q_{it})}$$

- (d) FISHER: este índice se define como la media geométrica de los índices de

Laspeyres y Paasche.

$$I_{F/0,t} = \sqrt{I_{L/0,t} \cdot I_{P/0,t}} = \sqrt{\frac{\sum_{i=1}^N p_{it}q_{i0}}{N} \cdot \frac{\sum_{i=1}^N p_{it}q_{it}}{N}}{\frac{\sum_{i=1}^N p_{i0}q_{i0}}{N} \cdot \frac{\sum_{i=1}^N p_{i0}q_{it}}{N}}$$

EJEMPLO 11.2 Se consideran los precios y cantidades vendidas de un conjunto de cuatro artículos para el período 1996-1998. Queremos calcular los índices de precios complejos ponderados definidos anteriormente para expresar la evolución del precio del conjunto de los cuatro artículos en 1998 respecto de 1996.

Artículos	Precios			Cantidades Vendidas		
	1996	1997	1998	1996	1997	1998
A	30	32	35	200	240	275
B	80	84	89	500	610	530
C	200	220	235	800	870	925
D	900	1100	1250	400	400	375

Los cálculos que se necesitan para obtener los citados índices aparecen en la siguiente tabla:

Artículos	$p_{i96}q_{i96}$	$p_{i98}q_{i96}$	$p_{i98}q_{i98}$	$p_{i96}q_{i98}$
A	6000	7000	9625	8250
B	40000	44500	47170	42400
C	160000	188000	217375	185000
D	360000	500000	468750	337500
Totales	566000	739500	742920	573150

$$I_{L/96,98} = \frac{\sum_{i=1}^4 p_{i98}q_{i96}}{\sum_{i=1}^4 p_{i96}q_{i96}} = \frac{739500}{566000} = 1,3065 \Rightarrow 130,65 \%$$

$$I_{P/96,98} = \frac{\sum_{i=1}^4 p_{i98}q_{i98}}{\sum_{i=1}^4 p_{i96}q_{i98}} = \frac{742920}{573150} = 1,2962 \Rightarrow 129,62 \%$$

$$I_{F/96,98} = \sqrt{I_{L/96,98} I_{P/96,98}} = \sqrt{1,3065 \times 1,2962} = 1,3013 \Rightarrow 130,13 \%$$

11.4. Propiedades de los índices complejos

Cuando se estudiaron los índices simples vimos que era recomendable que estos cumplieren una serie de propiedades. Así como los índices simples las cumplen en su gran mayoría, los complejos no cumplen algunas de ellas. En la siguiente tabla aparecen resumidas cuales de las propiedades estudiadas son verificadas, y cuales no, por los índices complejos vistos:

Propiedades	Índices					
	$I_{S/0,t}$	$I_{BD/0,t}$	$I_{L/0,t}$	$I_{P/0,t}$	$I_{MEB/0,t}$	$I_{F/0,t}$
Existencia	S	S	S	S	S	S
Identidad	S	S	S	S	S	S
Proporcionalidad	S	S	S	N(*)	N(*)	N(*)
Transitividad	N	S	N	N	N	N
Inversión	N	S	N	N	S	S

OBSERVACIÓN 11.2

- (a) El (*) indica que los correspondientes índices verifican algebraicamente la propiedad de proporcionalidad, aunque pueden realizarse objeciones de carácter económico.
- (b) El índice de Laspeyres es el más utilizado en los indicadores generales de precios en la mayoría de los países. Entre las razones que se tienen en cuenta para ello están:
- (i) Es el único índice complejo ponderado que realmente cumple la propiedad de proporcionalidad.
 - (ii) Cumple la propiedad de agregación, es decir, el índice de conjunto es igual al índice de Laspeyres de los índices de Laspeyres de cada subconjunto de componentes.
 - (iii) Es importante su significado económico.
 - (iv) Necesita datos menos actualizados. Su cálculo en cada período actual sólo requiere la actualización de los precios en dicho período si estamos calculando el índice de precios.

11.5. Enlaces y cambios de base

Un problema que se presenta en la elaboración de los números índices es la pérdida de representatividad a medida que el período actual se aleja del base. Por una parte, ocurre que las estructuras de consumo varían. Así, pueden aparecer artículos o servicios que hace tiempo eran impensables o desaparecer otros. Los consumos de los mismos también varían con el tiempo. Esto se puede mejorar

actualizando el período base cada cierto tiempo y así acercar el período base al actual.

Decidido que tenemos que modificar el período base, nos vamos a encontrar con dos series de índices referidas a distintos períodos bases. Se pretende que los índices de ambas series sean comparables. El procedimiento a partir del cual se construye una nueva serie de índices a partir de dos que estaban referidas a diferentes períodos base se llama enlace de series.

EJEMPLO 11.3 *Supongamos que en cierto país la serie de índices de precios referidos a 1980 son los que se dan en la siguiente tabla:*

Año	$I_{1980,t}$
1985	117
1986	122
1987	133
1988	150

Supongamos que a lo largo de ocho años se produce un cambio en la estructura socio-económica del país que obliga a un cambio de base en 1988 para el cálculo de los índices de precios. Esto obliga a cambiar productos, precios y ponderaciones, generando una nueva serie referida al año 1988.

Año	$I_{1988,t}$
1988	100
1989	152
1990	161

Si tenemos en cuenta la propiedad de transitividad, siendo t' la nueva base, tendremos:

$$I_{0,t} = I_{0,t'} \cdot I_{t',t} \Rightarrow I_{t',t} = \frac{I_{0,t}}{I_{0,t'}}$$

Sabemos que no todos los índices cumplen la propiedad transitiva, pero se actúa en la práctica como si la cumplieran ante la necesidad de efectuar enlaces de series. En estos casos los resultados son sólo aproximados.

Si lo que queremos es expresar los índices construidos según la base antigua en otros referidos a la nueva base (prolongar la serie hacia atrás), aplicando la propiedad de transitividad obtenemos

$$I_{1980,t} = I_{1980,1988} \cdot I_{1988,t} \Rightarrow I_{1988,t} = \frac{I_{1980,t}}{I_{1980,1988}} = \frac{I_{1980,t}}{1,5}$$

que indica que tendremos que dividir cada índice referido a la base antigua entre el índice enlace entre ambas series, $I_{1980,1988} = 1,5$.

<i>Año</i>	$I_{1980,t}$	$I_{1988,t}$
1985	117	78
1986	122	81.3
1987	133	88.6
1988	<u>150</u>	100
1989		152
1990		161

Pero también puede ser interesante tener los índices contruídos con la nueva base referidos a la base antigua (prolongar la serie hacia adelante). En este caso, teniendo de nuevo en cuenta la propiedad de transitividad

$$I_{1980,t} = I_{1980,1988} \cdot I_{1988,t} = 1,5 \cdot I_{1988,t}$$

tendremos que multiplicar cada índice referido a la base nueva por el índice enlace entre ambas series, $I_{1980,1988} = 1,5$.

<i>Año</i>	$I_{1980,t}$	$I_{1988,t}$
1985	117	
1986	122	
1987	133	
1988	<u>150</u>	100
1989	228	152
1990	241.5	161

11.6. Deflación de series estadísticas

Cuando tenemos que agregar un conjunto de bienes y servicios y estos gozan de gran heterogeneidad, es frecuente someterlos a un proceso de homogeneización a través de la obtención de su valor aplicando un sistema de precios. Es decir, multiplicaremos cantidades de bienes por sus respectivos precios lo que transformará nuestras cantidades físicas, generalmente heterogéneas, en valores homogéneos al estar expresados en la misma unidad monetaria. Esto hace que estos bienes y servicios sean sumables o agregables.

Los índices de valor nos permiten estudiar la evolución a lo largo del tiempo de la cuantificación monetaria de un conjunto de bienes y servicios. El valor de un conjunto de bienes y servicios en el período base viene dado por la expresión:

$$V_0 = \sum_{i=1}^N p_{i0} q_{i0}$$

y el valor en el período actual será:

$$V_t = \sum_{i=1}^N p_{it} q_{it}$$

Este valor se llama *nominal o a precios corrientes* porque los precios considerados son los del período de comparación. Si valoramos las cantidades a precios del período base, la serie se dirá que está expresada en *términos reales o a precios constantes* de ese período base.

En la tabla siguiente representamos dos series de valores, una a precios corrientes y otra a precios constantes:

Período t	Valores a precios corrientes (Serie de valores nominales)	Valores a precios constantes (Serie de valores reales)
0	$\sum_{i=1}^N p_{i0}q_{i0}$	$\sum_{i=1}^N p_{i0}q_{i0}$
1	$\sum_{i=1}^N p_{i1}q_{i1}$	$\sum_{i=1}^N p_{i0}q_{i1}$
$\vdots$	$\vdots$	$\vdots$
t	$\sum_{i=1}^N p_{it}q_{it}$	$\sum_{i=1}^N p_{i0}q_{it}$
$\vdots$	$\vdots$	$\vdots$
k	$\sum_{i=1}^N p_{ik}q_{ik}$	$\sum_{i=1}^N p_{i0}q_{ik}$

Para poder efectuar análisis comparativos de una serie de valor entre distintos períodos interesa eliminar el efecto de la inflación o deflación de sus respectivos precios. Es decir, hay que pasarla de precios corrientes o de cada período a precios constantes o del período que se considere como base. Esto es lo que se denomina *deflactar* la serie. Para ello debemos dividirla por el índice de precios que se considere más adecuado. El índice elegido recibe el nombre de *deflactor* de la serie.

Veamos qué efecto tienen como deflactores los índices de Laspeyres y Paasche:

(a) Utilizando el índice de Paasche como deflactor

$$\frac{V_t}{I_{P/0,t}} = \frac{\sum_{i=1}^N p_{it}q_{it}}{\sum_{i=1}^N p_{it}q_{it}} = \sum_{i=1}^N p_{i0}q_{it}$$

(b) Si utilizamos como deflactor el índice de Laspeyres

$$\frac{V_t}{I_{L/0,t}} = \frac{\sum_{i=1}^N p_{it}q_{it}}{\sum_{i=1}^N p_{it}q_{i0}} = \sum_{i=1}^N p_{i0}q_{i0} \cdot \frac{\sum_{i=1}^N p_{it}q_{it}}{\sum_{i=1}^N p_{it}q_{i0}} = V_0 \cdot I_{P_q/0,t}$$

OBSERVACIÓN 11.3 *El I_P es el deflactor más adecuado. Sin embargo se utiliza el I_L por ser el único disponible, en general, y porque su valor no difiere demasiado del proporcionado por el I_P .*

EJEMPLO 11.4 *El precio medio de una Ha. en la provincia de Cádiz, así como los índices de precios fueron, para los años 1980-1986, los siguientes:*

Años	Precios corrientes (en 10^5 pesetas)	$I_{80,t}$
1980	7.6	100.00
1981	8	106.36
1982	8.5	113.64
1983	9.1	120.00
1984	9.8	127.27
1985	11.4	134.55
1986	12.3	140.91

Queremos realizar el estudio de la evolución de la serie económica anterior a precios constantes del año 1980. Para ello si tenemos en cuenta que

$$\text{Precios constantes (del año 1980)} = \frac{\text{Precios corrientes (del año } t)}{I_{80,t}}$$

Años	Precios corrientes (en 10^5 pesetas)	$I_{80,t}$	Precios constantes (en 10^5 ptas. de 1980)
1980	7.6	100.00	7.60
1981	8	106.36	7.52
1982	8.5	113.64	7.48
1983	9.1	120.00	7.58
1984	9.8	127.27	7.70
1985	11.4	134.55	8.47
1986	12.3	140.91	8.73

EJEMPLO 11.5 *La siguiente tabla presenta las ganancias mensuales de cierto tipo de trabajador, así como los índices de precios con base 1996:*

Años	Ganancias mensuales (en euros)	$I_{96,t}$
1996	760	100
1997	820	107.3
1998	840	110.2
1999	850	111.3
2000	890	116.7

Queremos expresar la serie de salarios anterior a precios constantes del año 2000. Para ello tengamos en cuenta que

$$\text{Ganancia a precios constantes} = \frac{\text{Ganancia a precios corrientes del año } t}{I_{00,t}}$$

Necesitamos, por tanto, los índices de precios con base el año 2000. Para ello dividimos los índices originales entre $I_{96,00}$, es decir:

$$I_{00,t} = \frac{I_{96,t}}{I_{96,00}} = \frac{I_{96,t}}{1,167}$$

Años	$I_{00,t}$
1996	85.68
1997	91.94
1998	94.43
1999	95.37
2000	100

Por tanto, la serie de las ganancias a precios constantes del año 2000 será:

Años	Ganancias mensuales (en euros)
1996	887.02
1997	891.88
1998	889.54
1999	891.26
2000	890

11.7. Variación, repercusión y participación

En esta sección nos referiremos al caso del índice de Laspeyres. Nuestro objetivo es ver cómo contribuye cada uno de los i artículos o servicios que forman parte del cálculo del índice en la variación que experimenta éste de unos períodos a otros.

Definición 11.2 Dados los índices $I_{L/0,t}$ e $I_{L/0,t'}$, referidos a una misma base,

se denomina *variación* que experimenta el índice al pasar del período t al t' a

$$V_{t,t'} = I_{L/0,t'} - I_{L/0,t} = \frac{\sum_{i=1}^N p_{it'} q_{i0}}{\sum_{i=1}^N p_{i0} q_{i0}} - \frac{\sum_{i=1}^N p_{it} q_{i0}}{\sum_{i=1}^N p_{i0} q_{i0}} = \frac{\sum_{i=1}^N q_{i0} \cdot (p_{it'} - p_{it})}{\sum_{i=1}^N p_{i0} q_{i0}} = \sum_{i=1}^N R_i$$

Definición 11.3 A cada sumando R_i de la variación se le llama *repercusión* del i -ésimo componente del índice

$$R_i = \frac{q_{i0} \cdot (p_{it'} - p_{it})}{\sum_{i=1}^N p_{i0} q_{i0}}$$

OBSERVACIÓN 11.4 La repercusión del componente i puede expresarse en función del índice simple como:

$$R_i = \frac{p_{i0} \cdot \left(\frac{p_{it'}}{p_{i0}} - \frac{p_{it}}{p_{i0}} \right) \cdot q_{i0}}{\sum_{i=1}^N p_{i0} q_{i0}} = \frac{p_{i0} q_{i0} (I_{0,t'}^i - I_{0,t}^i)}{\sum_{i=1}^N p_{i0} q_{i0}} = \frac{\omega_i \cdot \Delta I^i}{\sum_{i=1}^N \omega_i} = W_i \cdot \Delta I^i$$

Definición 11.4 Se llama *participación* del i -ésimo componente del índice a la repercusión relativa expresada en porcentajes. Es decir:

$$P_i = \frac{R_i}{\sum_{i=1}^N R_i} \cdot 100$$

Definición 11.5 Se llama *tasa de variación* ó *variación relativa* de una cierta magnitud al pasar del período t al t' a:

$$V_{t,t'}^R = \frac{V_{t,t'}}{I_{L/0,t}} \cdot 100 = \frac{I_{L/0,t'} - I_{L/0,t}}{I_{L/0,t}} \cdot 100 = \left(\frac{I_{L/0,t'}}{I_{L/0,t}} - 1 \right) \cdot 100$$

EJEMPLO 11.6 Con los datos de la tabla siguiente, calcula el IPC de Agosto de 1987 y de Septiembre del mismo año. Calcula las repercusiones, participaciones

y tasa de variación de cada grupo del mes de Septiembre respecto del de Agosto.

Grupo(G_i)	AG 87	SEP 87	ω_i	ΔI^i	R_i	P_i	$V_{Ag, Sep}^R$
1. Alimentación	143,3	146,1	330	2,8	0,9240	72,57	1,9539
2. Vestido	146,6	147,6	87	1,0	0,0870	6,83	0,6821
3. Vivienda	126,8	126,9	186	0,1	0,0186	1,46	0,0788
4. Menaje	138,0	138,3	74	0,3	0,0222	1,74	0,2173
5. Medicina	132,2	132,6	24	0,4	0,0096	0,75	0,3025
6. Transportes	129,0	129,2	144	0,2	0,0288	2,26	0,1550
7. Cultura	134,9	136,3	70	1,4	0,0980	7,69	1,0378
8. Otros	148,0	149,0	85	1,0	0,0850	6,67	0,6756
			1000		1,2732	100	

Obsérvese que el IPC se calcula mediante la media aritmética ponderada de los ocho grupos principales:

$$\begin{aligned}
 IPC \text{ (Agosto 87)} &= \frac{\sum_{i=1}^8 \omega_i I(G_i)}{\sum_{i=1}^8 \omega_i} = \sum_{i=1}^8 W_i I(G_i) = \\
 &= (1,433) \cdot (0,330) + (1,466) \cdot (0,087) + (1,268) \cdot (0,186) + (1,380) \cdot (0,074) + \\
 &+ (1,322) \cdot (0,024) + (1,290) \cdot (0,144) + (1,349) \cdot (0,070) + (1,480) \cdot (0,085) = \\
 &= 1,376118 \Rightarrow 137,6118 \%
 \end{aligned}$$

$$\begin{aligned}
 IPC \text{ (Septiembre 87)} &= \frac{\sum_{i=1}^8 \omega_i I(G_i)}{\sum_{i=1}^8 \omega_i} = \sum_{i=1}^8 W_i I(G_i) = \\
 &= (1,461) \cdot (0,330) + (1,4476) \cdot (0,087) + (1,269) \cdot (0,186) + (1,383) \cdot (0,074) + \\
 &+ (1,326) \cdot (0,024) + (1,292) \cdot (0,144) + (1,363) \cdot (0,070) + (1,490) \cdot (0,085) = \\
 &= 1,388850 \Rightarrow 138,8850 \%
 \end{aligned}$$

Capítulo 12

Series temporales: análisis descriptivo

12.1. Concepto de serie temporal

El estudio de un fenómeno bajo una perspectiva temporal puede hacerse de forma diacrónica o longitudinal (observando su evolución a través del tiempo), o bien de forma sincrónica o seccional (referido a un momento concreto del tiempo).

En el presente capítulo tratamos de estudiar fenómenos económicos (consumo de las familias, los tipos de interés, gasto en un determinado grupo de gasto, el paro, etc.) a lo largo del tiempo. Con el estudio descriptivo de una serie temporal podemos describir históricamente el fenómeno de interés y predecir su comportamiento futuro.

Definición 12.1 *Llamamos serie temporal, cronológica, histórica o de tiempo a una sucesión de observaciones cuantitativas de un fenómeno, ordenadas en el tiempo. Una serie temporal se puede considerar como una variable bidimensional (t, Y_t) , en la que una de las componentes, la dependiente, es la magnitud que queremos analizar, mientras que la independiente es el tiempo.*

Todo análisis de una serie temporal debe iniciarse con una representación gráfica de la misma. Para ello utilizaremos un sistema de ejes cartesianos, representando en el de abscisas la variable tiempo, t , y los valores de la magnitud observada, Y_t , en el de ordenadas. Con esto se consigue el diagrama de dispersión de la distribución (t_i, y_{t_i}) . La unión mediante segmentos de sus puntos nos proporciona una "diagrama de sierra" del cual podremos extraer las conclusiones iniciales sobre el comportamiento de nuestra serie.

EJEMPLO 12.1 *En el siguiente ejemplo recogemos la serie temporal que nos da las ventas mensuales de vino de Jerez en la península, medidas en hectólitros (Fuente: Consejo Regulador)*

Meses	1992	1993	1994	1995	1996
Enero	12584.30	9221.21	6059.22	7187.26	9612.34
Febrero	15101.70	10215.79	11629.97	12146.97	6876.19
Marzo	15382.25	11031.00	12438.10	12566.55	10558.60
Abril	23138.75	18458.42	19830.65	21122.94	19327.43
Mayo	15080.00	15042.58	16308.59	17125.61	16347.49
Junio	11115.43	10844.27	7599.37	9404.73	8954.73
Julio	14419.57	11938.73	8866.79	8126.42	7138.00
Agosto	4141.55	7339.00	7377.25	6844.87	6585.36
Septiembre	10266.45	10985.69	10589.71	8874.79	11978.46
Octubre	14101.23	11088.72	10370.49	9226.27	10713.16
Noviembre	17281.77	14385.97	11427.58	13855.81	11169.65
Diciembre	21931.00	25156.46	18903.27	23627.86	20610.06

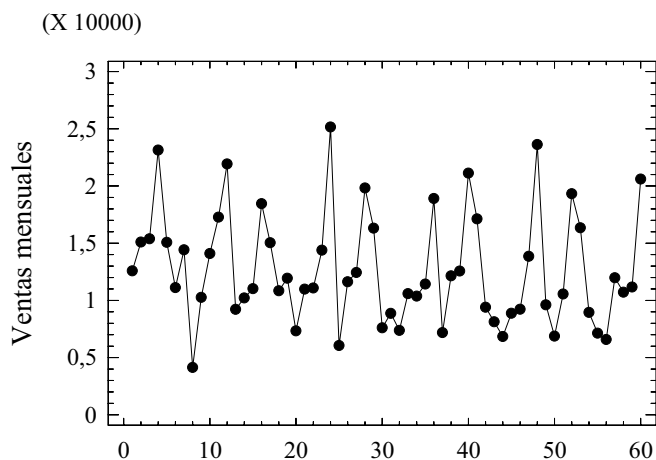


Figura 12.1: Serie temporal de las ventas mensuales de vino.

12.2. Descripción de una serie temporal

En el estudio clásico de las series temporales se supone que toda serie temporal está formada por cuatro componentes teóricas.

- (a) *Tendencia*: Expresa el carácter evolutivo de la serie a largo plazo. La denotaremos T_t .
- (b) *Estacional*: Expresa las fluctuaciones de la serie que se producen en un periodo igual o inferior a un año, y que se reproducen de manera recono-

cible en los diferentes años. Se deben, fundamentalmente, a efectos de la climatología sobre la actividad económica o a algunos hábitos sociales. Se representa por E_t .

- (c) *Cíclica*: Son las oscilaciones que se producen con un periodo superior al año, debidas a la alternancia de etapas de prosperidad y depresión. Se representan por C_t .
- (d) *Residual o irregular*: Movimientos originados por fenómenos imprevisibles, como huelgas, catástrofes, etc., que afectan a la variable en estudio de manera más o menos casual y no permanente, r_t .

¿Cómo se combinan las cuatro componentes teóricas para formar la serie que observamos? En el estudio clásico de las series temporales podemos considerar básicamente los modelos siguientes:

- (a) Modelo aditivo: $Y_t = T_t + E_t + C_t + r_t$.
- (b) Modelo multiplicativo: $Y_t = T_t \cdot E_t \cdot C_t + r_t$.

12.3. Análisis de la tendencia

Para analizar la tendencia o movimiento de una serie temporal a largo plazo existen varios métodos. Los más empleados son:

12.3.1. Método de ajuste analítico

Consiste en ajustar una función, que será la tendencia, a la nube de puntos que proporciona la serie temporal, aplicando el método de mínimos cuadrados.

Por tanto la ecuación de la tendencia será

$$Y_t(t) = f(t; a_1, a_2, \dots, a_n)$$

Recordemos que en primer lugar es conveniente realizar un análisis visual de la gráfica de la serie para elegir la forma de la función.

OBSERVACIÓN 12.1

- (a) Si en la serie temporal, la variable objeto de estudio viene medida de forma anual, ajustaremos la serie original (t_i, y_{t_i}) a una función adecuada.
- (b) Si se conoce el valor de la variable para períodos inferiores al año, representaremos con t_i el año i -ésimo y calcularemos las medias anuales de los valores de la variable Y_t correspondientes a dicho año, $\bar{y}_{t_i}$. Seguidamente ajustaremos la distribución $(t_i, \bar{y}_{t_i})$ a la función adecuada.

Los pasos a seguir son:

- (a) Calculamos las medias anuales $\bar{y}_{t_i}$ (si los datos vienen dados de forma anual, $\bar{y}_{t_i} = y_{t_i}$).
- (b) Se ajusta una función a la distribución anterior $(t_i, \bar{y}_{t_i})$.
- (i) Tendencia lineal: $T_t(t) = f(t; a, b) = a + bt$;

$$s.e.n. \hookrightarrow \begin{cases} \sum_j \bar{y}_{t_j} = aN + b \sum_i t_i \\ \sum_i \sum_j t_i \bar{y}_{t_j} = a \sum_i t_i + b \sum_i t_i^2 \end{cases}$$

- (ii) Tendencia hiperbólica: $T_t(t) = f(t; a, b) = a + b \cdot \frac{1}{t}$, a través del cambio $Z = \frac{1}{t}$.
- (iii) Tendencia potencial: $T_t(t) = f(t; a, b) = at^b$, mediante los cambios $U = \log(\bar{Y}_t)$ y $Z = \log(t)$.
- (iv) Tendencia exponencial: $T_t(t) = f(t; a, b) = ab^t$, haciendo uso del cambio $U = \log(\bar{Y}_t)$.

OBSERVACIÓN 12.2 *El sistema de ecuaciones normales se simplifica realizando el cambio de variable $t' = t - M_e(t)$, si el número de años es impar y siendo $M_e(t)$ el año mediano. Si el número de años es par, $t' = 2(t - M_e(t))$ es el cambio adecuado. Con los cambios propuestos se tiene que $\sum_i t'_i = 0$. Al final desharemos el cambio para obtener la ecuación de la tendencia en función de t .*

Ventajas del método de ajuste analítico

- (a) Expresa la tendencia a través de una función matemática con lo cual podemos realizar predicciones de cara al futuro.
- (b) Puede medirse la bondad del ajuste realizado mediante el coeficiente R^2 y así tener una medida de la fiabilidad de las predicciones.

EJEMPLO 12.2 *Vamos a calcular la tendencia lineal con los datos del Ejemplo 12.1. Como tenemos los valores de la variable para períodos inferiores al año, debemos previamente calcular las medias anuales $\bar{y}_{t_i}$, y realizar el ajuste de la distribución $(t_i, \bar{y}_{t_i})$ a la función elegida.*

- (a) Calculamos las medias anuales $\bar{y}_{t_i}$

$$\begin{aligned} \bar{y}_{1992} &= \frac{174544}{12}; \bar{y}_{1993} = \frac{155707,84}{12}; \bar{y}_{1994} = \frac{141400,99}{12}; \\ \bar{y}_{1995} &= \frac{150110,08}{12}; \bar{y}_{1996} = \frac{139871,47}{12} \end{aligned}$$

- (b) La tendencia lineal sería el resultado de realizar el ajuste lineal mínimo-cuadrático con los datos de la distribución $(t_i, \bar{y}_{t_i})$ dada en la siguiente tabla:

t_i	1992	1993	1994	1995	1996
$\bar{y}_{t_i}$	$\frac{174544}{12}$	$\frac{155707,84}{12}$	$\frac{141400,99}{12}$	$\frac{150110,08}{12}$	$\frac{139871,47}{12}$

Si tenemos en cuenta el comentario realizado en la Observación 12.2 y consideramos el cambio de variable $t' = t - 1994$, conseguimos anular alguno de los coeficientes en las ecuaciones del sistema de ecuaciones normales. Realizaremos este cambio, con lo cual, la distribución bidimensional considerada será:

t'_i	-2	-1	0	1	2
$\bar{y}_{t'_i}$	$\frac{174544}{12}$	$\frac{155707,84}{12}$	$\frac{141400,99}{12}$	$\frac{150110,08}{12}$	$\frac{139871,47}{12}$

Entonces

$$s.e.n. \hookrightarrow \begin{cases} \frac{761634,38}{12} = 5a + 0b \\ \frac{-74942,82}{12} = 0a + 10b \end{cases}$$

$$a = 12693,90633, \quad b = -624,5235$$

Luego la ecuación de la tendencia es:

$$T_t(t) = 12693,90633 - 624,5235(t - 1994) = 1257993,765 - 624,5235t$$

Para medir la fiabilidad de las predicciones que hagamos, calcularemos el coeficiente de determinación.

$$R^2 = \left(\frac{s_{t, \bar{Y}_t}}{s_t s_{\bar{Y}_t}} \right)^2 = \left(\frac{s_{t', \bar{Y}_{t'}}}{s_{t'} s_{\bar{Y}_{t'}}} \right)^2 = \frac{s_{t', \bar{Y}_{t'}}^2}{s_{t'}^2 s_{\bar{Y}_{t'}}^2} = 0,7159$$

donde

$$s_{t', \bar{Y}_{t'}} = \frac{-74942,82}{60} - 0 \cdot \frac{761634,38}{60} = -1249,047$$

$$s_{t'}^2 = \frac{10}{5} - 0^2 = 2$$

$$s_{\bar{Y}_{t'}}^2 = \frac{1,168018436 \times 10^{11}}{144 \cdot 5} - \left(\frac{761634,38}{60} \right)^2 = 1089524,756$$

Observamos que el coeficiente de determinación no alcanza el mínimo aceptable, con lo que el grado de fiabilidad de las posibles predicciones no será muy elevado.

12.3.2. Método de las medias móviles o método mecánico

Este método consiste en el suavizado de la serie dada, promediando sus observaciones con valores contiguos, anteriores y posteriores.

El procedimiento para calcular medias móviles de orden o tamaño h es el siguiente:

- La primera media móvil se obtiene calculando la media aritmética de las h primeras observaciones.
- Para calcular las siguientes, vamos excluyendo la primera observación del grupo anterior e incluyendo la posterior a la última tomada.
- El proceso se repite hasta que no se puedan formar más grupos que contengan h observaciones.

La tendencia será la línea quebrada que une las sucesivas medias móviles.

EJEMPLO 12.3 Usando medias móviles de orden 3,

t_i	y_{t_i}	Tendencia
t_1	y_{t_1}	
t_2	y_{t_2}	$\frac{y_{t_1} + y_{t_2} + y_{t_3}}{3} = \bar{y}_{t_2}$
t_3	y_{t_3}	$\frac{y_{t_2} + y_{t_3} + y_{t_4}}{3} = \bar{y}_{t_3}$
t_4	y_{t_4}	$\frac{y_{t_3} + y_{t_4} + y_{t_5}}{3} = \bar{y}_{t_4}$
t_5	y_{t_5}	$\frac{y_{t_4} + y_{t_5} + y_{t_6}}{3} = \bar{y}_{t_5}$
t_6	y_{t_6}	

La tendencia es la línea quebrada que une los puntos $(t_2, \bar{y}_{t_2})$, $(t_3, \bar{y}_{t_3})$, $(t_4, \bar{y}_{t_4})$ y $(t_5, \bar{y}_{t_5})$.

OBSERVACIÓN 12.3

- Existen observaciones para las que no se dispone de medias móviles.
- A mayor orden de las medias móviles, mayor suavizado, pero menor número de observaciones para cálculos posteriores.
- La elección del orden de las medias móviles no es fácil, y está ligado a las periodicidades de las fluctuaciones que se desean suavizar. Si los datos se refieren a períodos inferiores al año, se aconseja tomar como valor de h el número de dichos períodos. Cuando los datos de la serie son anuales, y por tanto no existe componente estacional, debemos tomar como orden el número de años que comprenda un ciclo.

- (d) Este método no permite efectuar predicciones ya que la tendencia no es una función matemática sino una serie amortiguada.
- (e) Cuando se calculen medias móviles de orden par, las observaciones no quedarán centradas en el tiempo. Por ello deberemos repetir el proceso a los promedios obtenidos inicialmente, utilizando el orden 2.

EJEMPLO 12.4 Usando medias móviles de orden 4,

t_i	y_{t_i}	$\bar{y}_{t_i}$	Tendencia
t_1	y_{t_1}	$\bar{y}_{t_3}$	$\bar{\bar{y}}_{t_3} = \frac{\bar{y}_{t_3} + \bar{y}_{t_4}}{2}$
t_2	y_{t_2}		
t_3	y_{t_3}	$\bar{y}_{t_4}$	$\bar{\bar{y}}_{t_4} = \frac{\bar{y}_{t_4} + \bar{y}_{t_5}}{2}$
t_4	y_{t_4}	$\bar{y}_{t_5}$	
t_5	y_{t_5}		
t_6	y_{t_6}		

La tendencia es la línea quebrada que une los puntos $(t_3, \bar{\bar{y}}_{t_3})$ y $(t_4, \bar{\bar{y}}_{t_4})$.

EJEMPLO 12.5 Durante el período 1975-1984, la inversión en instalaciones turísticas, expresada en millones de u.m., en cierto país fue la siguiente:

Años	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Inversión	600	800	750	400	350	500	1000	950	810	540

Suponiendo que la inversión turística se comporta cíclicamente con período de 4 años, en el cálculo de la tendencia por el método de las medias móviles, el valor de h sería 4. Calculemos dicha tendencia:

t_i	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
$\bar{\bar{y}}_{t_i}$			606,25	537,5	531,25	631,25	757,5	820		

Si el tamaño del ciclo fuese 5, las medias móviles deberían calcularse de orden 5.

t_i	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
$\bar{y}_{t_i}$			580	560	600	640	722	760		

12.4. Análisis de la estacionalidad

Recordemos que las variaciones estacionales se definían como aquellas oscilaciones de período igual o inferior a un año. En la gran mayoría de las series económicas las fluctuaciones debidas a la componente estacional pueden provocar una distorsión sobre la evolución real de la misma. Debemos, por tanto, identificar la componente estacional y a continuación eliminarla. A este procedimiento se le llama *desestacionalización*.

Existen diversos métodos para el cálculo de la componente estacional:

12.4.1. Método de las medias mensuales o método analítico

Supondremos como hipótesis que la tendencia es lineal.

Este método consta de los pasos siguientes:

- (a) Calculamos la tendencia.

Para ello se efectúa un ajuste lineal a los valores medios anuales, es decir, se determinan las medias anuales, $\bar{y}_{t_i}$, ajustándose a ellas una recta por el método de los mínimos cuadrados. Obtenida la ecuación de la tendencia

$$T_t(t) = a + bt$$

sabemos que la pendiente, b , es el incremento del valor medio de la variable debido al transcurso de un año. Por tanto, si p es el número de períodos estacionales ($p = 12$ para datos mensuales, $p = 4$ para datos trimestrales, etc.), $\frac{b}{p}$ nos proporciona el incremento del valor medio de la variable debido al transcurso de un período estacional. Es una estimación de la variación de la tendencia al pasar de un período estacional a otro.

- (b) Eliminamos la componente residual.

Para ello determinamos las medias estacionales (mensuales, trimestrales, etc.), $\bar{y}_{e_j}$. Se supone que al promediar oscilaciones residuales negativas se compensarán con oscilaciones residuales positivas.

- (c) Eliminamos la tendencia y la componente cíclica.

Obtenemos las medias estacionales corregidas eliminando de cada media estacional la parte correspondiente a la influencia de la tendencia, $\frac{b}{p}$, que se ha producido con el paso de un período a otro. También eliminamos la componente cíclica al considerarse que ésta se haya incluida en la tendencia.

$$\bar{y}'_{e_j} = \bar{y}_{e_j} - \frac{b(j-1)}{p}, \quad j = 1, 2, \dots, p$$

(d) Obtenemos la base.

A partir de las medias estacionales corregidas calculamos la media global corregida

$$\bar{y}' = \frac{\bar{y}'_{e_1} + \bar{y}'_{e_2} + \dots + \bar{y}'_{e_p}}{p}$$

(e) Calculamos las componentes estacionales

(i) Modelo aditivo: $e_j = \bar{y}'_{e_j} - \bar{y}'$

(ii) Modelo multiplicativo: $e_j = \frac{\bar{y}'_{e_j}}{\bar{y}'}$

OBSERVACIÓN 12.4 Las componentes estacionales, e_j , en el caso de que el modelo elegido sea el multiplicativo se suelen presentar multiplicadas por cien, y se llaman índices de variación estacional (I.V.E.).

EJEMPLO 12.6 La siguiente tabla recoge información sobre el consumo de materia prima por una empresa en los últimos seis años. Se quieren calcular las componentes estacionales.

	1998	1999	2000	2001	2002	2003
Trimestre 1	310	330	339	365	370	460
Trimestre 2	290	305	320	355	375	401
Trimestre 3	285	310	325	365	379	450
Trimestre 4	330	345	360	390	400	500

(a) Calculamos la tendencia, para lo que, previamente, calculamos las medias anuales.

$$\bar{y}_{1998} = \frac{1215}{4} = 303,75; \quad \bar{y}_{1999} = \frac{1290}{4} = 322,5; \quad \bar{y}_{2000} = \frac{1344}{4} = 336;$$

$$\bar{y}_{2001} = \frac{1475}{4} = 368,75; \quad \bar{y}_{2002} = \frac{1524}{4} = 381; \quad \bar{y}_{2003} = \frac{1811}{4} = 452,75$$

Queremos realizar un ajuste lineal para la distribución:

t_i	1998	1999	2000	2001	2002	2003
$\bar{y}_{t_i}$	303,75	322,5	336	368,75	381	452,75

Realizando el cambio de variable oportuno, $t' = 2(t - 2000,5)$, el ajuste lo realizaremos para la distribución:

t'_i	-5	-3	-1	1	3	5
$\bar{y}_{t_i}$	303,75	322,5	336	368,75	381	452,75

Así, el sistema de ecuaciones normales quedaría

$$s.e.n. \hookrightarrow \begin{cases} \frac{8659}{4} = 6a + 0b \\ \frac{3813}{4} = 0a + 70b \end{cases}$$

$$a = \frac{8659}{24} = 360,7916; \quad b = \frac{3813}{280} = 13,6178$$

Entonces

$$T_t(t) = 360,7916 + 13,6178(2(t - 2000,5)) = -54124,0262 + 27,2356t$$

- (b) Eliminamos la componente residual, para lo cual debemos calcular las medias trimestrales:

$$\bar{y}_{e_1} = \frac{2174}{6} = 362,3333; \quad \bar{y}_{e_2} = \frac{2046}{6} = 341;$$

$$\bar{y}_{e_3} = \frac{2114}{6} = 352,3333; \quad \bar{y}_{e_4} = \frac{2325}{6} = 387,5$$

- (c) Eliminamos la componente cíclica y la parte correspondiente a la influencia de la tendencia que se ha producido con el paso de un período a otro. Para ello hay que tener en cuenta el dato 27,2356 que es el valor de la pendiente en el ajuste lineal realizado. Entonces, las medias trimestrales corregidas son:

$$\bar{y}'_{e_1} = \bar{y}_{e_1} - \frac{27,2356(1-1)}{4} = 362,3333 - \frac{27,2356 \cdot 0}{4} = 362,3333$$

$$\bar{y}'_{e_2} = \bar{y}_{e_2} - \frac{27,2356(2-1)}{4} = 341 - \frac{27,2356 \cdot 1}{4} = 334,1911$$

$$\bar{y}'_{e_3} = \bar{y}_{e_3} - \frac{27,2356(3-1)}{4} = 352,3333 - \frac{27,2356 \cdot 2}{4} = 338,7155$$

$$\bar{y}'_{e_4} = \bar{y}_{e_4} - \frac{27,2356(4-1)}{4} = 387,5 - \frac{27,2356 \cdot 3}{4} = 367,0733$$

- (d) Calculamos la media global corregida

$$\bar{y}' = \frac{362,3333 + 334,1911 + 338,7155 + 367,0733}{4} = \frac{1402,3132}{4} = 350,5783$$

- (e) Tomando como base de comparación la media global corregida, obtenemos las componentes estacionales. El tratamiento es distinto si el modelo de comportamiento de la serie es aditivo o multiplicativo.

(i) Si el modelo fuese aditivo:

$$e_1 = \bar{y}'_{e_1} - \bar{y}' = 362,3333 - 350,5783 = 11,7550$$

$$e_2 = \bar{y}'_{e_2} - \bar{y}' = 334,1911 - 350,5783 = -16,3872$$

$$e_3 = \bar{y}'_{e_3} - \bar{y}' = 338,7155 - 350,5783 = -11,8628$$

$$e_4 = \bar{y}'_{e_4} - \bar{y}' = 367,0733 - 350,5783 = 16,4950$$

Si se supone que el incremento medio registrado en un trimestre considerado como "normal" es 0, entonces el consumo de materia por parte de la empresa se ve incrementado en 11,7550 unidades en los trimestres primeros y 16,4950 unidades en los trimestres cuartos de cada año, respecto de un trimestre medio. Por contra, en los trimestres segundo y tercero de cada año el consumo de materia descende en 16,3872 y 11,8628 unidades, respectivamente, respecto de un trimestre medio.

(ii) Si el modelo subyacente fuese multiplicativo:

$$e_1 = \frac{\bar{y}'_{e_1}}{\bar{y}'} = \frac{362,3333}{350,5783} = 1,0335 \Rightarrow 103,35 \%$$

$$e_2 = \frac{\bar{y}'_{e_2}}{\bar{y}'} = \frac{334,1911}{350,5783} = 0,9532 \Rightarrow 95,32 \%$$

$$e_3 = \frac{\bar{y}'_{e_3}}{\bar{y}'} = \frac{338,7155}{350,5783} = 0,9661 \Rightarrow 96,61 \%$$

$$e_4 = \frac{\bar{y}'_{e_4}}{\bar{y}'} = \frac{367,0733}{350,5783} = 1,0470 \Rightarrow 104,70 \%$$

El consumo de materia por parte de la empresa se ve incrementado en un 3,35% en los primeros trimestres y en un 4,7% en los trimestres cuartos de cada año, respecto de un trimestre medio. Por contra, en los trimestres segundo y tercero de cada año el consumo de materia descende en un 4,68% y en un 3,39%, respecto de un trimestre medio, respectivamente.

12.4.2. Método de las medias móviles o método mecánico

Este método consta de los siguientes pasos:

- Determinamos la tendencia calculando las medias móviles centradas en los períodos, $\bar{y}_{t_i}$ ó $\bar{\bar{y}}_{t_i}$.
- Eliminamos de forma conjunta la tendencia y la componente cíclica de los datos originales y_{t_i} .

- (i) Si el modelo es aditivo, por diferencia:

$$y_{t_i} - \bar{y}_{t_i} \quad \text{ó} \quad y_{t_i} - \bar{\bar{y}}_{t_i}$$

- (ii) Si el modelo es multiplicativo, por cociente:

$$\frac{y_{t_i}}{\bar{y}_{t_i}} \quad \text{ó} \quad \frac{y_{t_i}}{\bar{\bar{y}}_{t_i}}$$

- (c) Eliminamos la componente residual calculando los promedios de los valores obtenidos en el apartado (b) para cada período estacional:

- (i) Si el modelo es aditivo:

$$\bar{y}_{e_j} = \frac{1}{q_j} \sum_{i=1}^{q_j} (y_{t_i}^{(j)} - \bar{y}_{t_i}^{(j)}) \quad \text{ó} \quad \bar{y}_{e_j} = \frac{1}{q_j} \sum_{i=1}^{q_j} (y_{t_i}^{(j)} - \bar{\bar{y}}_{t_i}^{(j)})$$

donde q_j es el número de datos a promediar para el j -ésimo período estacional y los sumandos son los valores obtenidos en el apartado (b) para dicho período. Estos promedios son ya las diferentes componentes estacionales. Es decir, $e_j = \bar{y}_{e_j}$.

- (ii) Si el modelo es multiplicativo:

$$\bar{y}_{e_j} = \frac{1}{q_j} \sum_{i=1}^{q_j} \frac{y_{t_i}^{(j)}}{\bar{y}_{t_i}^{(j)}} \quad \text{ó} \quad \bar{y}_{e_j} = \frac{1}{q_j} \sum_{i=1}^{q_j} \frac{y_{t_i}^{(j)}}{\bar{\bar{y}}_{t_i}^{(j)}}$$

siendo q_j es el número de datos a promediar para el j -ésimo período estacional y los sumandos son los valores obtenidos en el apartado (b) para dicho período. Para obtener las componentes estacionales obtenemos la base, media aritmética de los valores anteriores, con la que efectuaremos las comparaciones:

$$\bar{y} = \frac{\bar{y}_{e_1} + \bar{y}_{e_2} + \dots + \bar{y}_{e_p}}{p}$$

Las componentes estacionales serán los *índices de variación estacional*

$$e_j = \frac{\bar{y}_{e_j}}{\bar{y}}.$$

EJEMPLO 12.7 Consideramos de nuevo los datos referentes al consumo de materia prima por parte de una empresa en los últimos seis años recogidos en el Ejemplo 12.6. Queremos calcular las componentes estacionales, empleando ahora el nuevo procedimiento explicado.

	1998	1999	2000	2001	2002	2003
<i>Trimestre 1</i>	310	330	339	365	370	460
<i>Trimestre 2</i>	290	305	320	355	375	401
<i>Trimestre 3</i>	285	310	325	365	379	450
<i>Trimestre 4</i>	330	345	360	390	400	500

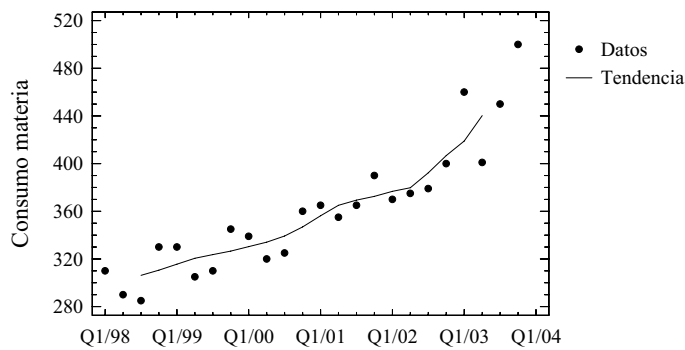


Figura 12.2: Serie original y tendencia

(a) *Determinamos la tendencia por el método de las medias móviles de orden igual a $p = 4$, calculando las medias móviles centradas, $\bar{\bar{y}}_{t_i}$.*

En un primer paso calculamos las medias móviles de tamaño 4 no centradas.

	1998	1999	2000	2001	2002	2003
Trimestre 1		318,75	332,25	361,25	378,50	427,75
Trimestre 2	303,75	322,50	336,00	368,75	381,00	452,75
Trimestre 3	308,75	324,75	342,50	370,00	403,50	
Trimestre 4	312,50	328,50	351,25	375,00	410,00	

Después calculamos las medias móviles de tamaño 4 centradas, $\bar{\bar{y}}_{t_i}$.

	1998	1999	2000	2001	2002	2003
Trimestre 1		315,625	330,375	356,250	376,750	418,875
Trimestre 2		320,625	334,125	365,000	379,750	440,250
Trimestre 3	306,250	323,625	339,250	369,375	392,250	
Trimestre 4	310,625	326,625	346,875	372,500	406,750	

(b) *Eliminamos la tendencia y la componente cíclica de los datos originales:*

(i) *Si el modelo es aditivo, haremos:*

$$y_{t_i} - \bar{\bar{y}}_{t_i}$$

	1998	1999	2000	2001	2002	2003
<i>T1</i>		14,375	8,625	8,750	−6,750	41,125
<i>T2</i>		−15,625	−14,125	−10,000	−4,750	−39,250
<i>T3</i>	−21,250	−13,625	−14,000	−4,375	−13,250	
<i>T4</i>	19,375	18,375	13,125	17,500	−6,750	

(ii) Si el modelo es multiplicativo, calcularemos:

$$\frac{y_{t_i}}{\bar{y}_{t_i}}$$

	1998	1999	2000	2001	2002	2003
<i>T1</i>		1,04554	1,02610	1,02456	0,98208	1,09817
<i>T2</i>		0,95126	0,95765	0,97260	0,98749	0,91084
<i>T3</i>	0,93061	0,95789	0,95799	0,98815	0,96622	
<i>T4</i>	1,06237	1,05625	1,03783	1,04697	0,98340	

(c) Eliminamos la componente residual calculando los promedios de los valores obtenidos en el apartado (b) para cada período estacional, es decir, para cada trimestre:

(i) Si el modelo es aditivo:

	Componente estacional: $e_j = \bar{y}_{e_j}$
Trimestre 1	13,225
Trimestre 2	−16,750
Trimestre 3	−13,300
Trimestre 4	12,325

Si se supone que el incremento medio registrado en un trimestre considerado como "normal" es 0, entonces el consumo de materia por parte de la empresa se ve incrementado en 13,225 unidades en los trimestres primeros y 12,325 unidades en los trimestres cuartos de cada año. Por contra, en los trimestres segundo y tercero de cada año el consumo de materia descende en 16,750 y 13,3 unidades, respectivamente.

(ii) Si el modelo es multiplicativo:

	Medias trimestrales: $\bar{y}_{e_j}$
Trimestre 1	1,03529
Trimestre 2	0,95596
Trimestre 3	0,96017
Trimestre 4	1,03736

A continuación obtenemos la base, es decir, la media de todos los valores anteriores:

$$\bar{y} = \frac{\bar{y}_{e_1} + \bar{y}_{e_2} + \bar{y}_{e_3} + \bar{y}_{e_4}}{p} = \frac{3,98878}{4} = 0,9971$$

Por último calcularemos las componentes estacionales

$$e_j = \frac{\bar{y}_{e_j}}{\bar{y}}$$

	Componente estacional: e_j
Trimestre 1	1,038202
Trimestre 2	0,958648
Trimestre 3	0,962870
Trimestre 4	1,040277

Multiplicadas por cien, obtenemos la expresión porcentual de más fácil interpretación:

$$e_1 : 103,8202\%; \quad e_2 : 95,8648\%; \quad e_3 : 96,2870\%; \quad e_4 : 104,0277\%$$

El consumo de materia por parte de la empresa se ve incrementado en un 3,8202% en los primeros trimestres y en un 4,0277% en los trimestres cuartos de cada año, respecto de un trimestre medio. Por contra, en los trimestres segundo y tercero de cada año el consumo de materia desciende en un 4,1352% y en un 3,713%, respecto de un trimestre medio, respectivamente.

Una vez obtenidas las componentes estacionales, podemos desestacionalizar la serie observada restándole a cada dato original de la correspondiente estación el valor de su componente estacional, si el modelo es aditivo, y dividiendo cada dato original entre la correspondiente componente estacional expresada en tantos por uno, en el caso multiplicativo.

(i) Si el modelo es aditivo:

	1998	1999	2000	2001	2002	2003
Trimestre 1	296,775	316,775	325,775	351,775	356,755	446,775
Trimestre 2	306,750	321,750	336,750	371,750	391,750	417,750
Trimestre 3	298,300	323,300	338,300	378,300	392,300	463,300
Trimestre 4	317,675	332,675	347,675	377,675	387,675	487,675

(ii) Si el modelo es multiplicativo:

	1998	1999	2000	2001	2002	2003
Trimestre 1	298,593	317,857	326,526	351,569	356,385	443,073
Trimestre 2	302,509	318,156	333,803	370,312	391,175	418,297
Trimestre 3	295,989	321,953	337,532	379,074	393,614	467,352
Trimestre 4	317,222	331,642	346,061	374,899	384,512	480,640

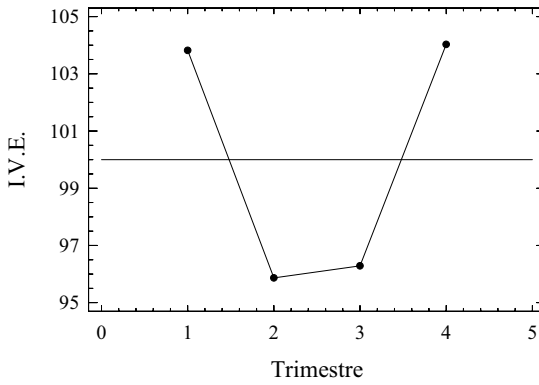


Figura 12.3: Indices de variación estacional

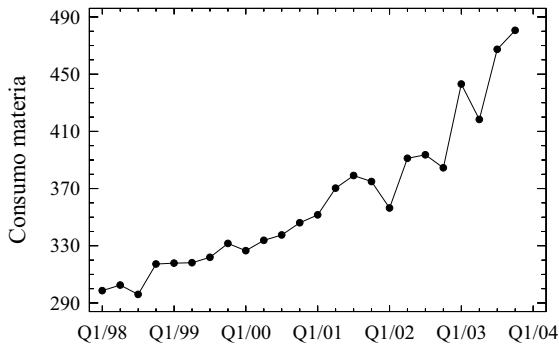


Figura 12.4: Serie desestacionalizada bajo el modelo multiplicativo

Cuestiones, problemas y recursos

Cuestiones

1. Los precios de dos artículos, A y B , en los años 1994 y 1995 fueron:

	1994	1995
A	12	12
B	25	28

- ¿Cómo ha variado el precio del artículo A entre los años 1994 y 1995? ¿Y el precio del conjunto formado por ambos artículos? Comenta qué tipo de índice has empleado en cada caso explicando por qué.
2. Cita un índice complejo ponderado y otro sin ponderar, explicando la diferencia entre ambos.
 3. ¿Por qué se construyen los índices complejos? ¿Qué relación guardan con los índices simples? Pon un ejemplo de un índice complejo donde señales esta relación.
 4. En la tabla siguiente se dan los índices de precios simples para tres mercancías. Se sabe, además, que el gasto realizado en cada mercancía en el año 1997 fue de 4, 6 y 10 unidades monetarias respectivamente.

Año	A	B	C
1997	100	100	100
1998	103	103	102
1999	107	109	108

- Calcula los índices de precios de Laspeyres tomando como año base el año 1997. ¿Podrían ser calculados los índices de Paasche?
5. ¿Qué son y qué utilidad tienen las ponderaciones?

6. Calcula e interpreta los índices de precios $I_{L/03,04}$, $I_{P/03,04}$, $I_{MEB/03,04}$ y $I_{F/03,04}$ con los siguientes datos:

	Precio-03	Precio-04	Cantidad-03	Cantidad-04
Artículo A	18	20	10	25
Artículo B	140	180	25	34

7. Un artículo subió en Diciembre respecto de Noviembre un 10 %, de Noviembre respecto de Octubre un 7 % y de Octubre respecto de Septiembre experimentó una bajada de un 5 %. Calcula el índice encadenado para ese artículo desde el mes de Diciembre respecto del de Septiembre.
8. Semejanzas y diferencias entre los índices de precios de Laspeyres y de Paasche.
9. El IPC ha tenido como período base diferentes años. ¿Por qué es aconsejable este cambio de base cada cierto tiempo?
10. La siguiente tabla recoge los índices de precios de consumo de Noviembre y Diciembre de 1999 en España para los ocho grandes grupos, así como las ponderaciones de los mismos (Fuente: Instituto Nacional de Estadística I.N.E.)

Grupos	ω_i	Noviembre 99	Diciembre 99
Alimentación	293	123.6	124.6
Vestido	115	121.1	121.2
Vivienda	103	133.7	134.3
Menaje	67	121.7	121.9
Medicina	31	128.9	129.0
Transporte	165	132.7	133.2
Cultura	73	124.6	124.6
Otros	153	135.2	135.8

Calcula e interpreta la participación del grupo Alimentación en la variación global del IPC entre los meses de Noviembre y Diciembre de 1999, sabiendo que esta variación fue de 0.5571.

11. En la siguiente tabla se recoge el sistema de ponderaciones empleado en España para la elaboración del IPC con base el año 1992.

Grupos	I	II	III	IV	V	VI	VII	VIII
ω_i	293	115	103	67	31	165	73	153

Siendo los grupos I a VIII, respectivamente, Alimentación, Vestido, Vivienda, Menaje, Medicina, Transporte, Cultura y Otros. ¿Qué significado tienen estos valores? ¿Por qué se emplean?

12. Enlaza las dos series de números índices con base el año 1994:

Año	1991	1992	1993	1994	Año	1994	1995	1996
$I_{90,t}$	105	107	120	130	$I_{94,t}$	100	120	125

13. Una nave industrial ha sido alquilada para su explotación al precio de 4500 euros en el año 1992. Si el índice de precios al consumo ha evolucionado según la siguiente tabla:

Año	1991	1992	1993	1994
IPC	100	106	111	120

¿Cuál será el precio del alquiler en 1994 si ese año se revisa el precio de acuerdo con el incremento del IPC?

14. Deflacta la siguiente serie de precios a pesetas constantes de 1990:

Año	1990	1991	1992	1993	1994	1995
Precios	1500	1850	1920	2400	2650	2800
IPC	100	105	115	120	125	130

15. La siguiente tabla contiene la evolución de los ingresos (en miles de euros) de una determinada agencia de viajes, así como los índices de precios al consumo con base 1992 en los mismos años :

Año	1997	1998	1999	2000
Ingresos	138	138	150	162
IPC	123	124.6	128.3	133.4

Calcula los ingresos totales, correspondientes al período 1997-2000, en unidades monetarias constantes del año 2000.

16. Interpreta la componente estacional $e_j = 1,85$ bajo un modelo multiplicativo y bajo un modelo aditivo.
17. Las componentes estacionales correspondientes a una serie temporal que mide la evolución de los precios, en miles de pesetas, de cierta mercancía a lo largo de varios años son:

	T1	T2	T3	T4
e_j	-1.25	6.25	-10	-1.875

Realiza una interpretación adecuada de las mismas.

18. Comenta brevemente las componentes de una serie temporal.

19. ¿Qué inconvenientes presenta el método de las medias móviles usado para el cálculo de la tendencia de una serie temporal?
20. Comenta qué es el orden de una media móvil, qué efecto tiene en relación a la serie original el cálculo de la tendencia por el método de las medias móviles y qué influencia tiene el cambio de orden de las medias móviles.
21. ¿En qué consiste la desestacionalización de una serie temporal? ¿Qué efecto produce sobre la serie original?

Problemas

1. Para medir la evolución de los precios en un conjunto formado por cinco bienes de consumo, se utilizó el índice de Laspeyres del cual se posee la siguiente información:

Año	$I_{80,t}$	$I_{92,t}$
1990	215	-
1991	225	-
1992	232	100
1993	-	103
1994	-	107

Con respecto a los años 1992 y 1995 se dispone de la siguiente información:

Bienes	Precios	Precios	Cantidades	Cantidades
	1992	1995	1992	1995
A	50	65	4500	4000
B	5000	5200	50	65
C	150	200	1250	1500
D	400	550	400	350
E	25	15	8000	12000

Se pide:

- La serie de índices de precios para el período 1990-1995 con base el año 1992.
 - Si una familia gasta 20000 u.m. en 1992 en bienes de tipo B, ¿cuánto le cuesta en 1995 comprar una cantidad equivalente?
 - Si una familia destinaba 70000 u.m. en 1992. ¿Cuánto deberá destinar a la adquisición de este conjunto de bienes en el año 1995 para que el poder adquisitivo sea el mismo que en el año 1992?
2. Una empresa cerámica produce tres calidades diferentes de baldosas, siendo el número de unidades producidas, en miles de m^2 y el coste unitario en cada modelo, en miles de u.m., durante los años 1993 y 1994, el que figura en la siguiente tabla:

Modelos	1993	1993	1994	1994
	Precio	Producción	Precio	Producción
A	1.1	20	1.3	18
B	2.2	12	2.8	15
C	3.1	5	3.5	5

- (a) Calcula e interpreta el índice de precios de Laspeyres del año 1994 respecto del año 1993.
- (b) Calcula el índice de precios de Paasche del año 1994 respecto del año 1993.
- (c) Si los índices de precios con base 1992 son:

Año	1994	1995	1996
$I_{92,t}$	105	110	118

y el precio de cada tipo de baldosa aumentó en el año 1995 de acuerdo con el índice de precios, calcula el precio de cada tipo de baldosa en dicho año.

3. La siguiente tabla muestra los salarios anuales de los nuevos diplomados de Gestión y Administración Pública, en miles de pesetas, y los IPC de los diversos años respecto de 1984:

Año	1985	1986	1987	1988	1989
Salarios	1697	1789	1812	2010	2215
IPC	107.6	109.6	113.6	118.3	123.5

- (a) ¿Cuáles son los salarios en términos reales en pesetas constantes del año 1989? ¿El salario ha ido siempre en aumento?
- (b) ¿Qué salario era necesario en 1989 para garantizar la misma renta real que en 1985? Compáralo con el que efectivamente recibe en 1989.
- (c) A un recién titulado se le ofrece un trabajo con una remuneración anual de 1912 miles de pesetas en Madrid y otro con remuneración de 1972 miles de pesetas en Sevilla. El coste de la vida en Madrid es de 126.5 y en Sevilla de 114.2. ¿En qué ciudad su sueldo es mayor en pesetas constantes?
- (d) Si en Barcelona el salario deflactado de un recién titulado es de 1550 miles de pesetas y el coste de la vida de Barcelona es un 2 % superior al de Madrid, en pesetas corrientes, ¿a cuánto asciende mi salario?
4. En un restaurante ha habido que reponer, durante el año 1999 el siguiente menaje: 100 cubiertos, 300 platos y 400 copas, y durante el año 2000 se repusieron 190, 250 y 500 unidades respectivamente. Los precios unitarios en el año 1996 de cada uno de estos artículos fue, respectivamente: 1000, 300 y 250 u.m. En la tabla siguiente se muestran los índices de precios para el período 1996-2000 con base el año 1992.

Año	1996	1997	1998	1999	2000
IPC	119,22	121,56	123,8	126,65	131

- (a) Calcula el precio unitario de cada artículo para los años 1999 y 2000, supuesto que estos han variado de acuerdo con el índice de precios.
- (b) Calcula el índice de precios de Fisher para conocer la evolución que ha sufrido el precio del conjunto formado por esos tres artículos en el año 2000 respecto del año 1999. Interpretalo.
- (c) Compara, en u.m. constantes, el gasto total realizado por el restaurante en el año 1999 y en el año 2000.
5. De un sistema de índices de precios de consumo se tiene la siguiente información estadística sobre los grupos de artículos que componen la cesta de la compra:

Grupos (G_i)	ω_i	$I_{93,95}(G_i)$	$I_{93,97}(G_i)$
I. Alimentación	40	119	130
II. Vestido	9	120	140
III. Vivienda	17	114	125
IV. Transporte	20	132	154
V. Otros	14	110	118

- (a) Completa la siguiente tabla:

Año	1993	1994	1995	1996	1997
IPC	100	110,69	119,58	127,93	

- (b) Calcula el incremento porcentual de los precios en el grupo Vivienda para el año 1997 respecto del año 1995. Calcula el incremento porcentual del IPC para el año 1997 respecto del año 1995.
- (c) Si el alquiler de un local se pactó en 1995 en 150000 u.m., ¿cuál será su valor actualizado en 1997 de acuerdo con la evolución de los precios dentro del grupo correspondiente? ¿Y si se pactó una revisión anual según el IPC? ¿Qué es más conveniente para el inquilino?
- (d) Estudia qué grupo o grupos han tenido una participación más importante en la variación del IPC del año 1997 con respecto del año 1995, interpretando adecuadamente tal participación.
6. Se dispone de la siguiente información sobre el Índice de Precios de Consumo Armonizado (IPCA) español con base el año 1996:

Grupos	ω_i	Enero 2000	Febrero 2000
Alimentos y bebidas	275	101,7	101,7
Bebidas alcohólicas y tabaco	32	120,9	123,9
Vestido y calzado	114	105,7	107,1
Vivienda	112	105,4	108,8
Menaje	65	104,8	106,1
Medicinas	8	105,6	106,9
Transporte	146	101,9	106,5
Comunicaciones	16	105,7	103,9
Ocio y cultura	69	105,0	109,0
Enseñanza	1	108,7	110,2
Hoteles y restaurantes	118	108,4	111,2
Otros	44	106,0	107,6
Total	1000		

Se pide:

- El IPCA general del mes de Enero. Interpretación.
 - La variación sufrida por el índice al pasar de Enero a Febrero a partir de las repercusiones.
 - Las participaciones de todos los grupos al pasar de Enero a Febrero y la interpretación del resultado del grupo más relevante.
 - Si el índice de Febrero del año 2000 con base el año 1996 es de 106.777, actualiza a pesetas de Febrero el precio de un artículo que en Enero de 2000 costaba 140000 pesetas.
7. Los beneficios totales semestrales, en miles de millones de pesetas, de un grupo de empresas vienen dados por la siguiente tabla:

Año	1990	1991	1992	1993	1994
Semestre 1	10	11	15	20	28
Semestre 2	14	17	19	22	32

- Calcula la tendencia lineal por el método del ajuste analítico.
 - Suponiendo un esquema aditivo, calcula e interpreta las componentes estacionales.
 - Desestacionaliza la serie comentando cómo lo haces.
8. La siguiente serie recoge el número de PYMES (Pequeñas y Medianas Empresas), que se crearon en Cádiz, durante el período 2000-2004:

	2000	2001	2002	2003	2004
Cuatrimstre 1	41	39	35	21	22
Cuatrimstre 2	37	33	27	15	16
Cuatrimstre 3	36	30	16	12	13

- Calcula la tendencia lineal por el método del ajuste analítico.
 - Calcula los índices de variación estacional por el método analítico e interpreta los mismos.
 - Desestacionaliza la serie comentando cómo lo haces.
9. Una región de un determinado país ha experimentado en cinco años las siguientes entradas de turistas por estación, medidos en millones:

Estación	1998	1999	2000	2001	2002
Primavera	3	4	4	5	6
Verano	11	14	16	19	20
Otoño	4	4	5	5	6
Invierno	2	2	3	3	4

- Calcula la tendencia, supuesta lineal, por el método del ajuste analítico (Se aconseja realizar un cambio de variable).
 - Realiza un pronóstico del número medio de turistas por estación para el año 2003.
 - Calcula los índices de variación estacional por el método del ajuste analítico. Realiza una interpretación adecuada de los mismos.
 - Desestacionaliza los datos correspondientes al año 2002.
10. La siguiente serie temporal muestra el número de botellas de Manzanilla Juncal vendidas en una tienda de barrio de Jerez en los últimos años:

	1997	1998	1999	2000	2001
Primer semestre	30	35	36	42	45
Segundo semestre	8	10	18	18	20

- Calcula la componente estacional por el método analítico, supuesto un modelo aditivo.
 - Haz lo mismo que en el anterior apartado pero por el método de las medias móviles.
11. El número de multas por estacionamiento indebido en una ciudad presenta la siguiente evolución para los últimos cuatro años:

Cuatrimestres	1996	1997	1998	1999
I	226	214	209	205
II	320	314	309	305
III	280	274	270	255

- (a) Calcula la tendencia de la serie por el método de las medias móviles.
- (b) Estudia las componentes estacionales supuesto un modelo aditivo.
- (c) Desestacionaliza la serie para el año 1999.

12. La siguiente tabla muestra el número de botellas de fino "Tío Mateo", vendidas en una tienda de barrio de Jerez en los últimos años:

Trimestres	1998	1999	2000	2001
I	35	36	42	45
II	10	18	18	20
III	40	30	36	40
IV	35	40	42	50

- (a) Calcula razonadamente e interpreta la componente estacional por el método de las medias móviles, supuesto un modelo aditivo.
- (b) Desestacionaliza la serie comentando cómo lo haces.

13. La siguiente tabla presenta las cifras de parados, expresadas en miles, para el período 1991-1994. Los datos han sido organizados por trimestres.

Trimestres	1991	1992	1993	1994
I	2700	2510	2420	2600
II	2555	2430	2390	2690
III	2470	2390	2480	2790
IV	2520	2425	2570	3050

- (a) Calcula la tendencia por el método de las medias móviles.
- (b) Calcula los índices de variación estacional. Interpretálos.
- (c) Desestacionaliza la serie para el año 1992.

14. La siguiente tabla refleja la evolución de los ingresos (en millones de pesetas) de una empresa industrial a lo largo de los cuatro años:

Trimestres	1996	1997	1998	1999
I	83	84	79	88
II	82	89	90	94
III	102	100	99	101
IV	108	112	124	142

- (a) Calcula e interpreta los índices de variación estacional, utilizando el método de las medias móviles.
 - (b) Dibuja la serie original y la serie desestacionalizada y haz un pequeño comentario.
15. En una determinada oficina se ha ido anotando cada mes el número de pasos telefónicos durante los años 1994-1998. Con el fin de simplificar la gran cantidad de datos anotados, se han calculado el número medio de pasos en cada semestre para los distintos años, obteniéndose:

	1994	1995	1996	1997	1998
Enero-Junio	171	176	180	183	187
Julio-Diciembre	180	189	191	193	196

- (a) Realiza la representación gráfica de la serie.
- (b) Calcula la tendencia por el método de las medias móviles.
- (c) Calcula e interpreta los índices de variación estacional, utilizando el método de las medias móviles. Desestacionaliza la serie.

Recursos

En el mercado nos encontramos con hojas de cálculo, como Excell, con la que fácilmente podemos calcular los diversos números índices que hemos dado.

También, en el año 2001, el INE ha puesto en marcha un sistema de almacenamiento y recuperación de la información estadística en Internet conocido como *INEbase*. Se trata de una puerta de entrada a la Estadística Oficial. Desde este portal se puede acceder al Banco de Datos del INE, TEMPUS, con más de 400 000 series temporales que van desde la esfera de la población a la contabilidad nacional, sociedad, empleo, etc.

Parte IV

PROBABILIDAD

Capítulo 13

Probabilidad

13.1. Experimentos aleatorios. Definiciones

Definición 13.1 *Llamaremos fenómenos causales o determinísticos a aquellos que presentan los mismos resultados si se realizan en idénticas condiciones. En ellos es posible conocer el resultado final conociendo el estado de partida y las condiciones de realización.*

EJEMPLO 13.1 *Cualquier experimento físico o químico.*

Definición 13.2 *Los fenómenos aleatorios son aquellos en los que no se puede predecir el resultado final incluso realizándose en las mismas condiciones.*

EJEMPLO 13.2 *Son ejemplos de experimentos aleatorios el lanzamiento de un dado equilibrado, la elección al azar de un número entre 0 y 1, consumo diario de agua de una ciudad, etc...*

Definición 13.3 *La Teoría de la Probabilidad estudia los métodos de análisis que son comunes en el tratamiento de los fenómenos aleatorios, cualquiera que sea el área en que éstos se presenten.*

Definición 13.4 *Se llama espacio muestral asociado a un experimento aleatorio al conjunto formado por todos los posibles resultados del experimento aleatorio.*

Definición 13.5 *Se llama punto muestral a cada elemento del espacio muestral.*

El espacio muestral suele representarse por Ω , y los puntos muestrales por ω .

EJEMPLO 13.3 *Consideremos el experimento aleatorio consistente en lanzar un dado equilibrado de seis caras al aire y observar el número de puntos que figuran en la cara superior. Su correspondiente espacio muestral será $\Omega = \{1, 2, 3, 4, 5, 6\}$.*

Definición 13.6 Se denomina suceso a todo subconjunto A del espacio muestral, ($A \subseteq \Omega$). Es un resultado en que se concreta el experimento.

Los sucesos suelen representarse por letras mayúsculas: $A, B, C, \dots$

EJEMPLO 13.4 En el lanzamiento de un dado, son sucesos:

$A = \text{"sacar puntuación par"} = \{2, 4, 6\};$

$B = \text{"sacar puntuación 2"} = \{2\}.$

Definición 13.7 Se llama espacio de sucesos, $\mathcal{S}$, al conjunto formado por todos los sucesos asociados al experimento aleatorio en cuestión.

Existen distintos tipos de sucesos:

- (a) *Suceso imposible* es aquel que no ocurre nunca. Se representa por $\emptyset$.
- (b) *Suceso seguro* es aquel que ocurre siempre. Se representa por Ω .
- (c) *Suceso elemental* es el formado por un sólo punto muestral.
- (d) *Suceso compuesto* es el formado por más de un punto muestral.

EJEMPLO 13.5 En el lanzamiento de un dado de seis caras sabemos que $\Omega = \{1, 2, 3, 4, 5, 6\}$, entonces el espacio de sucesos será:

$$\mathcal{S} = P(\Omega) = \{\emptyset, \{1\}, \dots, \{6\}, \{1, 2\}, \dots, \{5, 6\}, \{1, 2, 3\}, \dots, \{2, 3, 4, 5, 6\}, \Omega\}$$

Un suceso elemental es $B = \text{"obtener puntuación 2"} = \{2\}.$

Un suceso compuesto es $A = \text{"obtener puntuación par"} = \{2, 4, 6\}.$

13.2. Álgebra de sucesos. Propiedades

Hemos establecido una correspondencia entre sucesos y conjuntos. Vamos a recordar algunas operaciones y relaciones entre conjuntos, que ahora, serán de interés para trasladarlas a los sucesos.

Dados los sucesos A y B de un espacio muestral Ω , definimos:

Definición 13.8 El suceso complementario de A , que se denota por $\bar{A}$, es el suceso formado por todos los puntos muestrales que no pertenecen a A , es decir,

$$\bar{A} = \{\omega \in \Omega / \omega \notin A\}.$$

EJEMPLO 13.6 En el lanzamiento de un dado $\Rightarrow \Omega = \{1, 2, 3, 4, 5, 6\}.$

Dado $A = \text{"sacar par"} = \{2, 4, 6\} \Rightarrow \bar{A} = \text{"sacar impar"} = \{1, 3, 5\}$

Definición 13.9 La unión de los sucesos A y B , que se denota por $A \cup B$, es el suceso formado por todos los puntos muestrales que pertenecen al menos a uno de los sucesos,

$$A \cup B = \{\omega \in \Omega / \omega \in A \text{ o } \omega \in B\}$$

EJEMPLO 13.7 En el lanzamiento de un dado de seis caras, sean los sucesos A y B siguientes:

$A = \text{"sacar puntuación par"} = \{2, 4, 6\};$

$B = \text{"sacar puntuación mayor que 4"} = \{5, 6\}.$

Entonces

$$A \cup B = \text{"sacar puntuación par ó puntuación mayor que 4"} = \{2, 4, 5, 6\}$$

Definición 13.10 La intersección de los sucesos A y B , que denotamos por $A \cap B$, es el suceso formado por todos los puntos muestrales que pertenecen a ambos sucesos,

$$A \cap B = \{\omega \in \Omega / \omega \in A \text{ y } \omega \in B\}$$

EJEMPLO 13.8 En el lanzamiento de un dado de seis caras, sean los sucesos A y B siguientes:

$A = \text{"sacar puntuación par"} = \{2, 4, 6\};$

$B = \text{"sacar puntuación mayor que 4"} = \{5, 6\}.$

Entonces

$$A \cap B = \text{"sacar puntuación par y puntuación mayor que 4"} = \{6\}$$

Propiedades de las operaciones con sucesos

1. Conmutativas

$$(1.a) \quad A \cup B = B \cup A$$

$$(1.b) \quad A \cap B = B \cap A$$

2. Asociativas

$$(2.a) \quad A \cup (B \cup C) = (A \cup B) \cup C$$

$$(2.b) \quad A \cap (B \cap C) = (A \cap B) \cap C$$

3. Distributivas

$$(3.a) \quad A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$(3.b) \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

4. Involución

$$\overline{\overline{A}} = A$$

$$5.(5.a) \quad A \cup \emptyset = A$$

$$(5.b) \quad A \cap \emptyset = \emptyset$$

6. Idempotencia

(6.a) $A \cup A = A$

(6.b) $A \cap A = A$

7. Complementariedad

(7.a) $A \cup \bar{A} = \Omega$

(7.b) $A \cap \bar{A} = \emptyset$

8. (8.a) $A \cup \Omega = \Omega$

(8.b) $A \cap \Omega = A$

9. Leyes de De Morgan

(9.a) $\overline{A \cup B} = \bar{A} \cap \bar{B}$

(9.b) $\overline{A \cap B} = \bar{A} \cup \bar{B}$

Definición 13.11 Diremos que la familia $\mathcal{S}$ de subconjuntos de Ω tiene estructura de álgebra de Boole sobre Ω , si verifica las propiedades siguientes:

(a) $\Omega \in \mathcal{S}$

(b) $\forall A, B \in \mathcal{S} \Rightarrow A \cup B \in \mathcal{S}$

(c) $\forall A \in \mathcal{S} \Rightarrow \bar{A} \in \mathcal{S}$

OBSERVACIÓN 13.1 Si cambiamos la segunda propiedad, unión finita, por $\bigcup_i A_i \in \mathcal{S}$, unión numerable, diremos que la familia $\mathcal{S}$ tiene estructura de σ -álgebra sobre Ω .

Definición 13.12 A y B son sucesos incompatibles o mutuamente excluyentes, si la ocurrencia simultánea de ambos es imposible, es decir: $A \cap B = \emptyset$.

EJEMPLO 13.9 En el lanzamiento de un dado de seis caras, son incompatibles los sucesos:

$A =$ "sacar puntuación menor que 3" $= \{1, 2\}$ y $B =$ "sacar puntuación mayor que 4" $= \{5, 6\}$.

OBSERVACIÓN 13.2 Un suceso y su complementario son siempre sucesos incompatibles.

Luego existe una relación biunívoca entre conjuntos y sucesos:

Cálculo de Probabilidades	Teoría de Conjuntos
Suceso Seguro	Conjunto Universal
Suceso elemental	Punto del Conjunto Universal
Suceso	Subconjunto
Sucesos incompatibles	Conjuntos disjuntos
Unión de sucesos	Unión de conjuntos
Intersección de sucesos	Intersección de conjuntos
Suceso Imposible	Conjunto vacío
Suceso contrario	Conjunto complementario

13.3. Diversas concepciones de probabilidad

Queremos dar una medida de la ocurrencia de un suceso en un experimento aleatorio. Vamos a analizar los conceptos que se han desarrollado a lo largo de la historia con el propósito de analizar el concepto de probabilidad.

Dado un suceso, A , perteneciente al espacio de sucesos $\mathcal{S}$ asociado al espacio muestral Ω , la probabilidad trata de asignar a A una medida teórica de la ocurrencia de A .

(a) DEFINICIÓN CLÁSICA DE PROBABILIDAD Ó DE LAPLACE (1812)

Deben establecerse dos hipótesis necesarias:

- (i) El espacio muestral ha de ser finito, y
- (ii) Todos los sucesos elementales han de ser igualmente favorables

entonces se define la probabilidad del suceso A como

$$p(A) = \frac{\text{número de casos favorables a } A}{\text{número total de sucesos elementales posibles}} = \frac{\sharp(A)}{\sharp(\Omega)}$$

EJEMPLO 13.10 *En el lanzamiento de un dado de seis caras no cargado, consideremos $\Omega = \{1, 2, 3, 4, 5, 6\}$. Sea el suceso $A = \text{"sacar puntuación menor que 3"} = \{1, 2\}$, entonces:*

$$p(A) = \frac{\sharp(A)}{\sharp(\Omega)} = \frac{2}{6} = 0.\hat{3}$$

Inconvenientes

La regla de Laplace sólo es aplicable para el cálculo de probabilidades de sucesos que pertenezcan a espacios muestrales finitos y en los que los sucesos elementales puedan considerarse equiprobables.

Ventajas

Proporciona un método para calcular probabilidades.

(b) DEFINICIÓN FRECUENCIALISTA DE PROBABILIDAD O DE VON MISES (1919)

Esta definición está basada en el concepto de frecuencia relativa de ocurrencia de un suceso.

Si repetimos el experimento N veces, llamamos *frecuencia relativa* del suceso A , que denotamos por $f(A)$, al cociente entre el número de veces que éste se presenta y el total de pruebas. La frecuencia relativa no es más que una medida relativa y empírica de la ocurrencia de un suceso. Verifica las propiedades:

- (i) $0 \leq f(A) \leq 1$
- (ii) $f(\Omega) = 1$
- (iii) $f(A \cup B) = f(A) + f(B)$; si $A \cap B = \emptyset$

Es un hecho comprobado empíricamente que, la frecuencia relativa de un suceso tiende a estabilizarse cuando el número de pruebas aumenta. La definición frecuencialista de probabilidad se basa en este hecho, y asigna como probabilidad al suceso A el número:

$$p(A) = \lim_{N \rightarrow \infty} f(A) = \lim_{N \rightarrow \infty} \frac{n(A)}{N} = \lim_{N \rightarrow \infty} \frac{\text{frecuencia absoluta de } A}{\text{número total de pruebas}}$$

Estas conclusiones llevan el nombre de *Primera Ley de los Grandes Números*: Cuando el número de realizaciones de un experimento aleatorio crece, la frecuencia relativa del suceso asociado se va acercando cada vez más hacia un cierto valor. Este valor se llama *probabilidad del suceso*.

Ventajas

Permite calcular probabilidades cuando no se puede usar la definición clásica.

Inconvenientes

- (i) En esta definición de probabilidad no es conocida una expresión de $f(A)$ en términos de N .
- (ii) Las condiciones bajo las que se realiza el experimento pueden variar a lo largo del tiempo, y con ellas, las frecuencias relativas.
- (iii) En la práctica, no es posible repetir el experimento una infinidad de veces.

(c) DEFINICIÓN SUBJETIVISTA DE PROBABILIDAD O DE DE FINETTI (1931)

La mayoría de los sucesos son irrepetibles. La probabilidad personal o subjetiva es la valoración que hace un individuo de las posibilidades de obtener un resultado, es decir, la probabilidad se podría definir como las condiciones en que un individuo estaría dispuesto a apostar por ese suceso. Si un jugador cree que un caballo tiene un 50 % de posibilidad de ganar, hará la correspondiente apuesta.

(d) DEFINICIÓN AXIOMÁTICA O DE KOLMOGOROV (1933)

Está englobada dentro de la teoría de la medida. La probabilidad sería una medida de la incertidumbre, con propiedades similares a la las medidas de longitud, espacio, volumen, etc.

Dado el espacio de sucesos $\mathcal{S}$ asociado a un espacio muestral Ω , se define una medida de probabilidad, p , como una función:

$$p : \mathcal{S} \rightarrow [0, 1]$$

que verifique los siguientes axiomas:

Axioma 1: $p(A) \geq 0, \quad \forall A \in \mathcal{S}$

Axioma 2: $p(\Omega) = 1$

Axioma 3: $p\left(\bigcup_i A_i\right) = \sum_i p(A_i), \quad \forall A_i \in \mathcal{S}$ siempre que se verifique que $A_i \cap A_j = \emptyset$ si $i \neq j$

Ventajas

Desde el punto de vista matemático su definición como una medida de la incertidumbre es impecable.

Inconvenientes

No proporciona un método de cálculo de probabilidades, lo que hace que en la práctica tengamos que basarnos en las definiciones clásica y frecuentista.

13.4. Propiedades

Como consecuencia de los axiomas de Kolmogorov pueden deducirse una serie de propiedades de la función p , que enunciamos a continuación:

- (a) $p(\emptyset) = 0$.
- (b) $p\left(\bigcup_{i=1}^k A_i\right) = \sum_{i=1}^k p(A_i), \quad \forall A_i \in \mathcal{S} \text{ verificando que } A_i \cap A_j = \emptyset \text{ si } i \neq j.$
- (c) $p(\overline{A}) = 1 - p(A); \quad \forall A \in \mathcal{S}.$
- (d) Si $A \subseteq B \Rightarrow p(A) \leq p(B)$ con $A, B \in \mathcal{S}.$
- (e) $p(A \cup B) = p(A) + p(B) - p(A \cap B); \quad \text{con } A, B \in \mathcal{S}.$
- (f) $p\left(\bigcup_{i=1}^k A_i\right) \leq \sum_{i=1}^k p(A_i), \quad \forall A_i \in \mathcal{S}.$

Definición 13.13 Se denomina espacio probabilístico a la terna $(\Omega, \mathcal{S}, p)$, donde $\mathcal{S}$ es el espacio de sucesos asociado al espacio muestral Ω , y p es una medida de probabilidad.

Capítulo 14

Probabilidad condicionada

14.1. Probabilidad condicionada. Propiedades

En los ejemplos que hemos planteado hasta ahora, siempre hemos supuesto que cualquiera de los resultados puede ocurrir en el experimento. La incorporación de una información adicional, como por ejemplo, el conocimiento de la ocurrencia de otro suceso, puede hacer que determinados resultados no puedan ocurrir, con lo que el espacio muestral cambia y cambian las probabilidades.

EJEMPLO 14.1 *Supongamos el experimento consistente la extracción de una bola de un bolsa que contiene seis bolas numeradas del uno al seis y observar el resultado obtenido. El correspondiente espacio muestral es $\Omega = \{1, 2, 3, 4, 5, 6\}$, y la probabilidad inicial del suceso $A = \text{"sacar número primo"} = \{1, 2, 3, 5\}$ es:*

$$p(A) = \frac{4}{6} = \frac{2}{3}$$

Supongamos ahora que las bolas correspondientes a los números pares han sido introducidas en una bolsa de color rojo y las correspondientes a los impares en una de color amarillo. Seleccionamos al azar una de las dos bolsas resultando seleccionada la roja. Si a continuación extraemos una bola de dicha bolsa, ¿qué probabilidad hay de que la cifra obtenida sea número primo?

La información del color de la bolsa produce, en este caso, una reducción del espacio muestral a:

$$\Omega_{roja} = \Omega_{par} = \{2, 4, 6\}$$

con lo que,

$$p(A \text{ si se eligió bolsa roja}) = p(A/\text{puntuación par}) = \frac{1}{3}$$

Como vemos, en este caso, la información disponible ha hecho disminuir la probabilidad. Otras veces una información adicional aumenta dicha probabilidad.

Supongamos que el color de la bolsa seleccionada hubiese sido amarilla, entonces:

$$\Omega_{\text{amarilla}} = \Omega_{\text{impar}} = \{1, 3, 5\}$$

y, por tanto,

$$p(A \text{ si se eligió bolsa amarilla}) = p(A/\text{puntuación impar}) = 1$$

Definición 14.1 Cuando consideremos la probabilidad de ocurrencia de un suceso A perteneciente a un espacio de sucesos sabiendo que ha acontecido otro suceso B , diremos que estamos calculando la probabilidad de A condicionada a B . Lo denotamos por $p(A/B)$, donde A es el suceso condicionado y B es el suceso condicionante.

En el ejemplo anterior podemos expresar la probabilidad de obtener número primo, habiendo obtenido cifra par como:

$$p(A/\text{puntuación par}) = \frac{1}{3} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{p(A \cap \text{puntuación par})}{p(\text{puntuación par})}$$

y, la probabilidad de obtener número primo, habiendo obtenido cifra impar como:

$$p(A/\text{puntuación impar}) = \frac{3}{3} = \frac{\frac{3}{6}}{\frac{3}{6}} = \frac{p(A \cap \text{puntuación impar})}{p(\text{puntuación impar})}$$

Definición 14.2 Sea (Ω, S, p) un espacio probabilístico y B un suceso de S con probabilidad no nula, $p(B) > 0$. Sea A un suceso cualquiera de S , llamaremos probabilidad del suceso A condicionada porque haya acontecido otro suceso B o, sencillamente, probabilidad de A condicionada por B , al cociente

$$p(A/B) = \frac{p(A \cap B)}{p(B)}$$

EJEMPLO 14.2 Se lanza tres veces una moneda equilibrada y se sabe que en la primera tirada salió cara. ¿Cuál es la probabilidad de obtener al menos dos caras? Llamaremos:

A = "obtener al menos dos caras en los tres lanzamientos";

B = "obtener cara en el primer lanzamiento".

Queremos calcular la probabilidad:

$$p(A/B) = \frac{p(A \cap B)}{p(B)} = \frac{\frac{3}{8}}{\frac{4}{8}} = \frac{3}{4} = 0,75$$

Sea el espacio probabilístico (Ω, S, p) y B es un suceso de S con $p(B) > 0$. La función $p(*|B)$, que a cada suceso A de S le asocia el número $p(A/B)$, es una probabilidad, y por tanto, se verifican las propiedades demostradas en el tema anterior para esta función.

$$(a) \quad p(\emptyset/B) = 0$$

$$(b) \quad p\left(\bigcup_{i=1}^k A_i/B\right) = \sum_{i=1}^k p(A_i/B), \quad \forall A_i \in S \text{ tal que } A_i \cap A_j = \emptyset \text{ si } i \neq j$$

$$(c) \quad p(\bar{A}/B) = 1 - p(A/B); \quad \forall A \in S$$

$$(d) \quad \text{Si } A_1 \subseteq A_2 \Rightarrow p(A_1/B) \leq p(A_2/B) \text{ con } A_1, A_2 \in S$$

$$(e) \quad p(A_1 \cup A_2/B) = p(A_1/B) + p(A_2/B) - p(A_1 \cap A_2/B); \text{ con } A_1, A_2 \in S$$

$$(f) \quad p\left(\bigcup_{i=1}^k A_i/B\right) \leq \sum_{i=1}^k p(A_i/B), \quad \forall A_i \in S$$

14.2. Teorema del producto

Teorema 14.1 Sean $A_1, A_2, \dots, A_n \in S$ tales que $p(A_1 \cap A_2 \cap \dots \cap A_{n-1}) \neq 0$ entonces

$$p(A_1 \cap A_2 \cap \dots \cap A_n) = p(A_1) \cdot p(A_2/A_1) \cdot p(A_3/A_1 \cap A_2) \cdots p(A_n/A_1 \cap A_2 \cap \dots \cap A_{n-1})$$

EJEMPLO 14.3 De un lote de doce artículos, de los cuales cuatro son defectuosos, se toman tres artículos escogidos al azar uno tras otro sin reemplazamiento. Calcula la probabilidad de que los tres artículos sean no defectuosos.

Sean los sucesos

$A_1 = "$ el primer artículo seleccionado es no defectuoso $"$;

$A_2 = "$ el segundo artículo seleccionado es no defectuoso $"$;

$A_3 = "$ el tercer artículo seleccionado es no defectuoso $"$.

Entonces:

$$p(A_1 \cap A_2 \cap A_3) = p(A_1) \cdot p(A_2/A_1) \cdot p(A_3/A_1 \cap A_2) = \frac{8}{12} \cdot \frac{7}{11} \cdot \frac{6}{10} = \frac{14}{55}$$

14.3. Sucesos dependientes e independientes

Estudiemos el siguiente ejemplo:

EJEMPLO 14.4 Consideremos el experimento consistente en lanzar un dado no cargado y sean los sucesos A y B siguientes:

$$A = "$$
 obtener cifra mayor que 2 $" = \{3, 4, 5, 6\} \Rightarrow p(A) = \frac{4}{6}$

$$B = \text{"obtener cifra par"} = \{2, 4, 6\} \Rightarrow p(B) = \frac{3}{6}$$

$$A \cap B = \text{"obtener cifra par mayor que 2"} = \{4, 6\} \Rightarrow p(A \cap B) = \frac{2}{6}$$

entonces

$$p(A/B) = \frac{p(A \cap B)}{p(B)} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3} = p(A)$$

$$p(B/A) = \frac{p(B \cap A)}{p(A)} = \frac{\frac{2}{6}}{\frac{4}{6}} = \frac{2}{4} = p(B)$$

Lo que observamos es que la información suministrada por el suceso condicionante resulta indiferente en cuanto a la probabilidad de ocurrencia del suceso condicionado. En este caso los sucesos A y B se dirán independientes.

Definición 14.3 Sea el espacio probabilístico (Ω, S, p) y sean A y B sucesos de S con $p(B) > 0$. Diremos que los sucesos A y B son independientes si se verifica que

$$p(A/B) = p(A)$$

O dicho de otra forma:

Definición 14.4 Diremos que dos sucesos A y B son independientes si y solo si se verifica que:

$$p(A \cap B) = p(A) \cdot p(B)$$

OBSERVACIÓN 14.1

(a) Sea un espacio probabilístico (Ω, S, p) y sean dos sucesos A y B de S . Si A y B son independientes, también lo son A y $\overline{B}$ (lo mismo puede afirmarse del par de sucesos $\overline{A}$ y B , y del par $\overline{A}$ y $\overline{B}$).

Demostración: El suceso A puede expresarse como la unión de los sucesos incompatibles:

$$A = (A \cap B) \cup (A \cap \overline{B}) \Rightarrow p(A) = p(A \cap B) + p(A \cap \overline{B})$$

Como A y B son independientes $\Rightarrow p(A \cap B) = p(A) \cdot p(B)$, entonces se tiene:

$$p(A \cap \overline{B}) = p(A) - p(A \cap B) = p(A) - p(A) \cdot p(B) = p(A) \cdot p(\overline{B})$$

Por tanto,

$$p(A \cap \overline{B}) = p(A) \cdot p(\overline{B}) \Rightarrow A \text{ y } \overline{B} \text{ son independientes}$$

- (b) No existe ninguna implicación entre los conceptos de incompatibilidad e independencia.

EJEMPLO 14.5 Sea el experimento consistente en el lanzamiento de un dado no cargado. Sean los sucesos $A = \{1, 2\}$, $B = \{2, 4, 6\}$ y $C = \{3, 4\}$. Podemos comprobar que:

- (i) A y B son independientes pero no incompatibles.

$A \cap B = \{2\} \Rightarrow A$ y B no son incompatibles.

$$p(A \cap B) = \frac{1}{6} = \frac{2}{6} \cdot \frac{3}{6} = p(A) \cdot p(B) \Rightarrow A \text{ y } B \text{ son independientes.}$$

- (ii) A y C son incompatibles pero no independientes.

$A \cap C = \emptyset \Rightarrow A$ y C son incompatibles

$$p(A \cap C) = 0 \neq \frac{2}{6} \cdot \frac{2}{6} = p(A) \cdot p(C) \Rightarrow A \text{ y } C \text{ no son independientes.}$$

14.4. Teorema de la probabilidad total

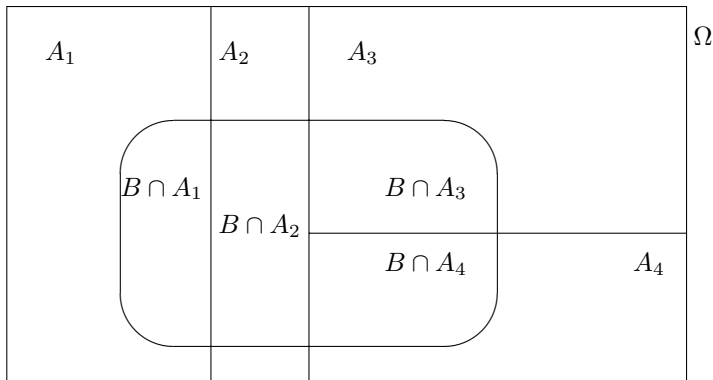
Teorema 14.2 Dado un espacio probabilístico (Ω, S, p) , si $A_1, A_2, \dots, A_n \in S$ es una colección de sucesos mutuamente excluyentes, todos con probabilidades no nulas, y tales que $\Omega = \bigcup_{i=1}^n A_i$, se verifica para todo $B \in S$:

$$p(B) = \sum_{i=1}^n p(B/A_i) \cdot p(A_i)$$

Demostración:

Podemos expresar el suceso B como unión de sucesos mutuamente excluyentes,

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^n A_i \right) = \bigcup_{i=1}^n (B \cap A_i)$$



Por tanto, teniendo en cuenta la propiedad (b) que verifica la probabilidad, así como el teorema del producto, se tiene:

$$p(B) = p\left(\bigcup_{i=1}^n (B \cap A_i)\right) = \sum_{i=1}^n p(B \cap A_i) = \sum_{i=1}^n p(B/A_i) \cdot p(A_i)$$

EJEMPLO 14.6 *Se quiere estudiar la situación laboral de los trabajadores en cuatro sectores de la economía. Las probabilidades de seleccionar a una persona en paro en cada uno de los cuatro sectores son, respectivamente, 0,06; 0,03; 0,02 y 0,1. Si el 40 % de las personas pertenecen al primer sector, y el resto se divide en parte iguales entre los otros tres, ¿cuál es la probabilidad de que una persona seleccionada al azar esté en paro?*

Consideremos los sucesos

A_1 = "la persona seleccionada pertenece al primer sector económico";

A_2 = "la persona seleccionada pertenece al segundo sector económico";

A_3 = "la persona seleccionada pertenece al tercer sector económico";

A_4 = "la persona seleccionada pertenece al cuarto sector económico";

B = "la persona seleccionada está en paro".

De estos sucesos conocemos

$$p(A_1) = 0,4; \quad p(A_2) = 0,2; \quad p(A_3) = 0,2; \quad p(A_4) = 0,2;$$

$$p(B/A_1) = 0,06; \quad p(B/A_2) = 0,03; \quad p(B/A_3) = 0,02; \quad y \quad p(B/A_4) = 0,1.$$

Entonces

$$\begin{aligned} p(B) &= p(B/A_1) \cdot p(A_1) + p(B/A_2) \cdot p(A_2) + p(B/A_3) \cdot p(A_3) + p(B/A_4) \cdot p(A_4) = \\ &= 0,06 \cdot 0,4 + 0,03 \cdot 0,2 + 0,02 \cdot 0,2 + 0,1 \cdot 0,2 = 0,054 \end{aligned}$$

14.5. Teorema de Bayes

Teorema 14.3 *Dado un espacio probabilístico (Ω, S, p) , si $A_1, A_2, \dots, A_n \in S$ es una colección de sucesos mutuamente excluyentes, todos con probabilidades no nulas, y tales que $\Omega = \bigcup_{i=1}^n A_i$, se verifica para todo $B \in S$:*

$$p(A_j/B) = \frac{p(B/A_j) \cdot p(A_j)}{\sum_{i=1}^n p(B/A_i) \cdot p(A_i)}, \quad \text{con } j = 1, 2, \dots, n.$$

Demostración: Teniendo en cuenta la definición de probabilidad condicionada y el teorema de la probabilidad total, se obtiene

$$p(A_j/B) = \frac{p(A_j \cap B)}{p(B)} = \frac{p(B/A_j) \cdot p(A_j)}{\sum_{i=1}^n p(B/A_i) \cdot p(A_i)}, \quad \text{con } j = 1, 2, \dots, n.$$

A las probabilidades $p(A_j)$ se les llama *probabilidades a priori*, y son las probabilidades iniciales que tenemos de los sucesos A_j . Ante una determinada evidencia experimental, B , corregimos el grado de creencia de las A_j obteniendo unas nuevas probabilidades, $p(A_j/B)$, llamadas *probabilidades a posteriori*, a través de las *verosimilitudes*, $p(B/A_j)$.

EJEMPLO 14.7 Con el enunciado del Ejemplo 14.6, supongamos que la persona seleccionada está parada, ¿cuál es la probabilidad de que pertenezca al primer sector económico?

Entonces la probabilidad pedida será:

$$p(A_1/B) = \frac{p(B/A_1) \cdot p(A_1)}{\sum_{i=1}^4 p(B/A_i) \cdot p(A_i)} = \frac{0,06 \cdot 0,4}{0,054} = 0,4\hat{}$$

Cuestiones, problemas y recursos

Cuestiones

1. La probabilidad de que un hombre viva 10 años más es $1/4$ y la probabilidad de que viva su esposa 10 años más es de $1/3$. Calcula la probabilidad de que dentro de 10 años:
(a) Ambos estén vivos; (b) Al menos uno de los dos esté vivo.
2. Un experimento consiste en el lanzamiento independiente de un dado y una moneda (no cargados). Describe el espacio muestral y calcula la probabilidad de obtener cara y puntuación par.
3. ¿Son independientes los sucesos A y B si $p(A) = 0,15$; $p(B) = 0,82$ y $p(A \cap B) = 0,11$?
4. Definición de sucesos independientes e incompatibles.
5. Se considera el experimento aleatorio consistente en extraer una carta de una baraja española (40 cartas). Sean los sucesos siguientes: A = "sacar una carta de oros ", B = "sacar una carta de copas ", C = "sacar un rey " y D = "sacar una figura ". Estudia si pueden considerarse independientes y si pueden considerarse incompatibles los pares de sucesos:
(a) A y B ; (b) A y C ; (c) C y D .
6. Pon un ejemplo razonado de un experimento aleatorio y de un suceso del mismo al cual podamos aplicar la regla de Laplace para el cálculo de su probabilidad. (Plantea el experimento, espacio muestral, suceso al que se le va a calcular la probabilidad, así como las condiciones de aplicación de la regla de Laplace).
7. Diferencias entre los experimentos aleatorios y determinísticos.
8. Diferencia entre sucesos simples y compuestos.

9. Ventajas e inconvenientes de la definición laplaciana de probabilidad.
10. Ventajas e inconvenientes de la definición frecuencalista de probabilidad.
11. Enuncia la definición de Kolmogorov de probabilidad.
12. Definición de espacio muestral. Ejemplo.
13. Sea el espacio muestral $\Omega = A \cup B \cup C$ siendo A , B y C incompatibles dos a dos. Deduce razonadamente si puede darse que $p(A) = 0,6$; $p(B) = 0,2$ y $p(C) = 0,3$.
14. Considérense los sucesos A y B tales que $p(A) = 0,6$ y $p(B) = 0,7$. Deduce razonadamente si son o no incompatibles.
15. Sean A y B dos sucesos tales que $p(A) = 0,3$; $p(A \cup B) = 0,8$ y $p(B) = p$. ¿Para qué valor de p son A y B independientes?
16. Consideremos los sucesos A , B , C , y D , tales que A , B y C son incompatibles entre sí y $\Omega = A \cup B \cup C$. Sabiendo que $p(D/A) = 0,3$; $p(A) = 0,6$ y $p(B) = p(D) = 0,3$, calcula $p(C)$ y $p(A/D)$.
17. Dados los sucesos A y B , pertenecientes al mismo espacio de sucesos, con $p(A) = 0,2$; $p(\overline{B}) = 0,5$ y $p(A \cup B) = 0,6$. Calcula $p(A/B)$ y razona si A y B pueden considerarse sucesos independientes. ¿Pueden considerarse incompatibles?
18. Sean A y B sucesos pertenecientes al mismo espacio de sucesos, con $p(\overline{A} \cup \overline{B}) = 0,7$, $p(A) = 0,4$ y $p(A/B) = 0,5$. Calcula $p(B)$. ¿Pueden considerarse independientes los sucesos A y B ?
19. Sean A y B sucesos del mismo espacio de sucesos tales que con $p(A) = 0,7$; $p(B) = 0,6$ y $p(\overline{A} \cup \overline{B}) = 0,6$. Calcula $p(A \cap B)$, $p(A \cup B)$ y $p(\overline{A} \cap \overline{B})$.
20. Sean A y B sucesos pertenecientes al mismo espacio de sucesos tales que $p(A) = 0,2$; $p(B) = 0,5$ y $p(A \cap B) = 0,4$. Razona si son coherentes los datos que se dan.

Problemas

1. Un estudiante sabe que hay siete temas que pueden preguntarle en un examen de Estadística Administrativa. Sólo dispone de tiempo para estudiar un tema, y lo selecciona al azar. Supongamos que la probabilidad de aprobar si el tema seleccionado es el que le preguntan es 0.9, mientras que si le preguntan uno de los otros seis temas la probabilidad de aprobar es de sólo 0.18.
 - (a) ¿Cuál es la probabilidad de que dicho estudiante apruebe?
 - (b) Si el estudiante ha aprobado el examen, ¿cuál es la probabilidad de que el tema del examen fuera el seleccionado?
 - (c) ¿Cuál es la probabilidad de no aprobar y que le caiga el tema del examen que él seleccionó?
2. El hotel NH AVENIDA de Jerez consigue automóviles para sus huéspedes de tres agencias de alquiler, el 20 % de la agencia Avis, el 40 % de la agencia Europcar y el otro 40 % de la agencia Hertz. Si el 14 % de los automóviles de Avis, el 4 % de Europcar y el 8 % de Hertz necesitan una puesta a punto. Calcular la probabilidad de que:
 - (a) Se entregue a los huéspedes un coche que necesite una puesta a punto.
 - (b) Si se entrega a los huéspedes un coche que necesita una puesta a punto, éste sea de la agencia Hertz.
 - (c) Si se entrega a los huéspedes un coche que no necesita una puesta a punto, éste sea de la agencia Avis.
3. Tres encuestadores A, B y C realizan 100, 110 y 120 encuestas al día, respectivamente, resultando que A escribe datos incorrectamente en el 12 % de las encuestas que realiza, que B los escribe incorrectamente en el 18 % de sus encuestas, y C en el 22 %.
 - (a) Calcular la probabilidad de escoger una encuesta incorrecta al azar.
 - (b) ¿Y la de escoger una encuesta correcta?
 - (c) Si se toma al azar una encuesta correcta de esas 330. ¿Cuál es la probabilidad de que haya sido de A?
4. El médico del Jerez Industrial sabe que el 40 % de los partidos de su equipo se juegan sobre campos de tierra. También sabe que la probabilidad de que un jugador sufra una lesión de rodilla si juega en este tipo de superficie se duplica al caso de que se juegue sobre hierba. Si la probabilidad de sufrir una lesión de rodilla sobre hierba es de 0.21, calcula la probabilidad de que:
 - (a) Un jugador de fútbol seleccionado aleatoriamente se lesione en una rodilla.

- (b) Un jugador de fútbol que sufre una lesión de rodilla se haya lesionado jugando sobre hierba.
 - (c) Un jugador de fútbol que no se haya lesionado haya jugado sobre hierba.
5. En la Diplomatura de Gestión y Administración Pública se sabe que el 40 % de los alumnos se ha matriculado en asignaturas de primer curso, el 30 % en segundo curso y el resto en tercer curso. También se sabe que el 25 % de los alumnos de primero abandonan sus estudios. Este porcentaje cae al 10 % en segundo y al 2 % en tercero. Con esto datos:
- (a) Calcula la probabilidad de que un alumno elegido al azar abandone sus estudios.
 - (b) Un determinado alumno ha abandonado sus estudios. Calcular la probabilidad de que sea de segundo curso.
 - (c) Calcula la probabilidad de que un alumno haya abandonado sus estudios y sea de tercero.
6. En un campus universitario existen tres carreras jurídico-sociales. Se sabe que el 50 % de los alumnos cursan estudios de Derecho, el 30 % de Empresa y el 20 % de Economía. Es conocido que el porcentaje de finalización de estudios es del 20 %, 10 % y 5 %, respectivamente. Elegido un estudiante al azar, hállese la probabilidad de que:
- (a) Haya acabado la carrera.
 - (b) Sea de Derecho y no haya acabado la carrera.
 - (c) Si el alumno ha acabado la carrera, que haya estudiado Economía.
 - (d) O sea de Empresa o no haya acabado la carrera.
7. Una compañía de seguros de automóviles clasifica los conductores en tres clases: A (alto riesgo), B (riesgo medio) y C (bajo riesgo). La clase A constituye el 30 % de los conductores que suscriben el seguro con la compañía; la probabilidad de que uno de esos conductores tenga algún accidente en un año es 0,1. Los correspondientes datos para la clase B son 50 % y 0,03, mientras que para la clase C son 20 % y 0,01.
- (a) Un determinado cliente contrata una póliza de seguros y en el primer año tiene un accidente. ¿Cuál es la probabilidad de que este cliente esté suscrito a la clase C ?
 - (b) Un determinado cliente contrata una póliza de seguros y en el primer año no tiene accidentes. ¿Cuál es la probabilidad de que este cliente esté suscrito a la clase A ?

8. Un fabricante estadounidense de bicicletas de carreras tiene naves de distribución en tres ciudades: San Luis, con el 50 % de la producción, Toledo con el 20 % y Chicago. Estableció controles de calidad en sus naves de distribución de las tres ciudades para inspeccionar las bicicletas y determinar si tenían algún defecto. El control de calidad comprobaba cada bicicleta y especificaba el resultado de la inspección como defectuosa o no defectuosa. Los resultados al inspeccionar las bicicletas fueron:

Ciudad	Resultado de la inspección
San Luis	40 % defectuosas
Toledo	25 % defectuosas
Chicago	10 % defectuosas

Seleccionada una bicicleta al azar, calcula la probabilidad de que:

- (a) Proceda de Toledo y sea defectuosa.
 - (b) Sea no defectuosa.
 - (c) Proceda de San Luis o no sea defectuosa.
 - (d) Si la bicicleta es defectuosa, proceda de Chicago.
9. En un sistema de alarma la probabilidad de que se produzca un peligro es 0,1. Si éste se produce, la probabilidad de que la alarma funcione es 0,95. La probabilidad de que funcione la alarma sin haber existido peligro es 0,03. Calcula la probabilidad de que:
- (a) Habiendo funcionado la alarma, no haya existido peligro.
 - (b) Haya existido peligro y la alarma no funcione.
 - (c) No habiendo funcionado la alarma, haya existido un peligro.

Recursos

Utilizando la tecla Random de cualquier calculadora, podemos generar números aleatorios y simular el lanzamiento de dados y monedas. ¿Te atreverías a simular el funcionamiento de una ruleta?

Parte V

MODELOS DE DISTRIBUCIONES

Capítulo 15

Variables aleatorias

15.1. Variable aleatoria: concepto

EJEMPLO 15.1 *Realicemos el experimento consistente en lanzar una moneda no cargada dos veces. Su espacio muestral será:*

$$\Omega = \{(c, c), (c, +), (+, c), (+, +)\}$$

donde todos los puntos muestrales son equiprobables.

Nos fijaremos en una determinada característica numérica del experimento, como por ejemplo, $X = \text{"número de caras obtenidas en los dos lanzamientos"}$.

Podemos considerar X como una aplicación que asocia a cada resultado del espacio muestral un valor numérico

$$\begin{array}{ll} X : \Omega & \longrightarrow \mathbb{R} \\ (c, c) & \longrightarrow 2 \\ (c, +) & \longrightarrow 1 \\ (+, c) & \longrightarrow 1 \\ (+, +) & \longrightarrow 0 \end{array}$$

Además, cada uno de estos valores se toma con una cierta probabilidad inducida por la aleatoriedad del fenómeno al que está asociado. Así, podemos escribir, por ejemplo:

$$\begin{aligned} p[X = 0] &= p[(+, +)] = \frac{1}{4} \\ p[X = 1] &= p[(c, +) \cup (+, c)] = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \\ p[X = 2] &= p[(c, c)] = \frac{1}{4} \end{aligned}$$

También pueden obtenerse probabilidades de intervalos:

$$p[X < 2] = p[(+, +) \cup (+, c) \cup (c, +)] = \frac{3}{4}$$

Definición 15.1 Sea $(\Omega, \mathcal{S}, P)$ un espacio probabilístico, se denomina *variable aleatoria (v.a.)* a una aplicación:

$$\begin{aligned} X : \quad \Omega &\longrightarrow \mathbb{R} \\ w \in \Omega &\longrightarrow X(w) \in \mathbb{R} \end{aligned}$$

OBSERVACIÓN 15.1 En términos matemáticos precisos es necesario exigir alguna condición adicional a la función para que se pueda considerar como variable aleatoria. Sin embargo, en los casos que vamos a manejar esta condición se verifica de forma elemental. La noción de variable aleatoria es la de una función que asigna un valor numérico a cada suceso elemental. De este modo trasladamos la probabilidad definida sobre sucesos a subconjuntos de la recta real.

Como podemos recordar las variables estadísticas median una característica numérica observada en los elementos de una población, contabilizándose los valores observados en términos de sus frecuencias relativas. En cambio, los posibles valores de una variable aleatoria vendrán contabilizados en términos de probabilidades. Usaremos la función de distribución para asignar probabilidades a ciertos sucesos definidos en términos de variables aleatorias.

15.2. Función de distribución. Propiedades

Definición 15.2 Se denomina *función de distribución (f.d.D.)* de una variable aleatoria X a la función F_X definida como sigue:

$$F_X : \mathbb{R} \longrightarrow [0, 1]$$

$$F_X(x) = p[X \leq x], \quad \forall x \in \mathbb{R}.$$

La función de distribución de la variable aleatoria X describe la acumulación de probabilidad por la variable a lo largo de la recta real. Tiene su antecedente en la frecuencia relativa acumulada.

Una función de distribución verifica las siguientes propiedades:

- (a) $F_X(-\infty) = \lim_{x \rightarrow -\infty} F_X(x) = 0.$
- (b) $F_X(+\infty) = \lim_{x \rightarrow +\infty} F_X(x) = 1.$
- (c) F_X es monótona no decreciente.
- (d) F_X es continua por la derecha, es decir, $F_X(x) = \lim_{h \rightarrow 0^+} F_X(x+h).$
- (e) $p[a < X \leq b] = F_X(b) - F_X(a),$ con $a, b \in \mathbb{R}.$

EJEMPLO 15.2 Consideremos el Ejemplo 15.1, y sea X "número de caras obtenidas en dos lanzamientos de la moneda". Sus posibles valores son $X = \{0, 1, 2\}$. Calculemos primero F_X en los posibles valores de X :

$$\begin{aligned} F_X(0) &= p[X \leq 0] = p[X = 0] = \frac{1}{4} \\ F_X(1) &= p[X \leq 1] = p[X = 0] + p[X = 1] = \frac{3}{4} \\ F_X(2) &= p[X \leq 2] = p[X = 0] + p[X = 1] + p[X = 2] = 1 \end{aligned}$$

Observemos que F_X está definida en todo el conjunto de los números reales, por tanto:

$$F_X(x) = \begin{cases} 0, & \text{si } x < 0 \\ 1/4, & \text{si } 0 \leq x < 1 \\ 3/4, & \text{si } 1 \leq x < 2 \\ 1, & \text{si } x \geq 2 \end{cases}$$

15.3. Variables aleatorias discretas

Sea $(\Omega, \mathcal{S}, P)$ un espacio probabilístico:

Definición 15.3 Una variable aleatoria X es discreta si el conjunto de valores que puede tomar X con probabilidad no nula es discreto (finito ó infinito numerable).

Si la variable es discreta y toma pocos valores distintos, podemos dar esos valores con sus probabilidades de forma explícita.

Pero si la variable es discreta con muchos valores diferentes o continua, debemos apoyarnos en funciones, como la función masa de probabilidad.

Definición 15.4 La función de masa de probabilidad ó función de probabilidad de una variable aleatoria discreta X que toma los valores $x_1, x_2, \dots, x_n, \dots$ con probabilidades no nulas es una función

$$p_X : \mathbb{R} \rightarrow [0, 1]$$

definida por:

$$p_X(x) = \begin{cases} p[X = x_k], & \text{si } x = x_k, \quad k = 1, 2, \dots, n, \dots \\ 0, & \text{en otro caso.} \end{cases}$$

Sea X una variable aleatoria discreta que toma los valores $x_1, x_2, \dots, x_n, \dots$ entonces se verifican las siguientes propiedades:

(a) $0 \leq p_X(x_k) \leq 1$ para todo k .

(b) $\sum_k p_X(x_k) = 1$.

$$(c) F_X(x) = p[X \leq x] = \sum_{x_k \leq x} p_X(x_k).$$

$$(d) p_X(x_k) = F_X(x_k) - F_X(x_{k-1}).$$

EJEMPLO 15.3 *Vamos a calcular la función de probabilidad y la función de distribución de la variable X = "número total de bolas blancas obtenidas al hacer dos extracciones con reemplazamiento en una urna compuesta por una bola blanca, una roja y una negra".*

El espacio muestral correspondiente es:

$$\Omega = \{(b, b), (b, r), (b, n), (r, b), (r, r), (r, n), (n, b), (n, r), (n, n)\}$$

por lo que el conjunto de los posibles valores de la variable es $X = \{0, 1, 2\}$, con probabilidades

$$p[X = 0] = \frac{4}{9}, \quad p[X = 1] = \frac{4}{9}, \quad p[X = 2] = \frac{1}{9},$$

que determinan la función de probabilidad.

La función de distribución vendrá dada por:

$$F_X(x) = \begin{cases} 0, & \text{si } x < 0 \\ 4/9, & \text{si } 0 \leq x < 1 \\ 8/9, & \text{si } 1 \leq x < 2 \\ 1, & \text{si } x \geq 2 \end{cases}$$

OBSERVACIÓN 15.2 *Observemos que la función de distribución de una variable discreta es escalonada y los saltos de la función de distribución se producen justamente en los valores que toma la variable y son de amplitud igual a las probabilidades con que los toma. Es decir,*

$$p[X = 0] = F_X(0) - F_X(0^-) = \frac{4}{9} - 0 = \frac{4}{9}$$

$$p[X = 1] = F_X(1) - F_X(0) = \frac{8}{9} - \frac{4}{9} = \frac{4}{9}$$

$$p[X = 2] = F_X(2) - F_X(1) = 1 - \frac{8}{9} = \frac{1}{9}$$

15.4. Variables aleatorias continuas

Sea $(\Omega, \mathcal{S}, P)$ un espacio probabilístico:

Definición 15.5 *Una variable aleatoria X con función de distribución F_X se dice que es continua, si existe una función $f_X(x) \geq 0$ tal que:*

$$F_X(x) = p[X \leq x] = \int_{-\infty}^x f_X(t) dt, \quad \forall x \in \mathbb{R}.$$

A $f_x(x)$ se le denomina *función de densidad (f.d.d.)* de la variable aleatoria continua X .

Asociadas a la función de densidad tenemos las siguientes propiedades:

- (a) $\int_{-\infty}^{+\infty} f_x(t) dt = 1, (F_x(+\infty) = 1).$
- (b) $f_x(x) = F'_x(x)$, es decir, la f.d.d. puede obtenerse a través de la f.d.D.
- (c) $p[X = a] = 0, \forall a \in \mathbb{R}.$ (Como F_x es continua, no tiene saltos).
- (d)

$$p[X \leq b] = p[X < b] = F_x(b) = \int_{-\infty}^b f_x(t) dt.$$

$$p[X > a] = p[X \geq a] = 1 - p[X \leq a] = 1 - F_x(a) = \int_a^{+\infty} f_x(t) dt.$$

$$\begin{aligned} p[a < X \leq b] &= p[a \leq X < b] = p[a < X < b] = p[a \leq X \leq b] = \\ &= F_x(b) - F_x(a) = \int_a^b f_x(t) dt \end{aligned}$$

OBSERVACIÓN 15.3 Una función de densidad de probabilidad representa una aproximación muy útil para calcular probabilidades partiendo de un histograma, permite:

- (a) Sustituir la tabla completa de valores de la distribución de frecuencias por una ecuación matemática $f_x(x)$.
- (b) Tratar de reflejar no una muestra concreta, sino la estructura de distribución de los valores de la variable a largo plazo. Los valores más altos de la función de densidad se corresponden con zonas en las que es más probable que aparezca resultados de un fenómeno aleatorio.
- (c) Es más operativa que el histograma ya que con la función de densidad puedo calcular probabilidades.

EJEMPLO 15.4 Sea una v.a. X con función de densidad:

$$f_x(x) = \begin{cases} kx^2, & \text{si } 0 < x < 2 \\ 0, & \text{en otro caso} \end{cases}$$

(a) Calcula el valor de k que hace que f_x sea función de densidad.

$$\begin{aligned} 1 &= \int_{-\infty}^{+\infty} f_x(t) dt = \int_{-\infty}^0 f_x(t) dt + \int_0^2 f_x(t) dt + \int_2^{+\infty} f_x(t) dt = \\ &= \int_0^2 kt^2 dt = k \frac{2^3}{3} \Rightarrow k = \frac{3}{8} \end{aligned}$$

(b) Obténgase la correspondiente función de distribución.

Primer Tramo: ($x < 0$), $F_x(x) = \int_{-\infty}^x f_x(t) dt = \int_{-\infty}^x 0 dt = 0$

Segundo Tramo: ($0 \leq x < 2$),

$$F_x(x) = \int_{-\infty}^x f_x(t) dt = \int_{-\infty}^0 0 dt + \int_0^x \frac{3}{8} t^2 dt = \frac{x^3}{8}$$

Tercer Tramo: ($x \geq 2$),

$$F_x(x) = \int_{-\infty}^x f_x(t) dt = \int_{-\infty}^0 0 dt + \int_0^2 \frac{3}{8} t^2 dt + \int_2^x 0 dt = 1$$

Por tanto, la función de distribución es

$$F_x(x) = \begin{cases} 0, & \text{si } x < 0 \\ \frac{x^3}{8}, & \text{si } 0 \leq x < 2 \\ 1, & \text{si } x \geq 2 \end{cases}$$

(c) Calcula $p[1 < X < 1,5]$ y $p[X > 0,5]$.

$$p(1 < X < 1,5) = F_x(1,5) - F_x(1) = \frac{(1,5)^3}{8} - \frac{1}{8} = \frac{3,375}{8} = 0,29688$$

$$p(X > 0,5) = 1 - p(X \leq 0,5) = 1 - F_x(0,5) = 1 - \frac{(0,5)^3}{8} = 0,9843$$

EJEMPLO 15.5 Sea la variable X "Hora en la que el reloj deja de funcionar por agotarse la pila", que presenta la siguiente función de distribución:

$$F_x(x) = \begin{cases} 0, & \text{si } x < 0 \\ \frac{x}{24}, & \text{si } 0 \leq x < 24 \\ 1, & \text{si } x \geq 24 \end{cases}$$

Queremos obtener la correspondiente función de densidad. Para ello aplicaremos la propiedad $f_x(x) = F'_x(x)$. Entonces

$$f_x(x) = \begin{cases} \frac{1}{24}, & \text{si } 0 < x < 24 \\ 0, & \text{en otro caso} \end{cases}$$

15.5. Características de las variables aleatorias

15.5.1. Introducción

EJEMPLO 15.6 Una empresa contabiliza el número de accidentes laborales diarios. Los datos del último mes fueron:

Número de accidentes	0	1	2	3	4
Número de días	10	12	5	2	1

Considerando la variable estadística X = "número de accidentes diarios", puede considerarse la distribución de frecuencias:

x_i	0	1	2	3	4
n_i	10	12	5	2	1
f_i	1/3	2/5	1/6	1/15	1/30

Para dicha distribución podemos calcular una serie de coeficientes como por ejemplo $\bar{x}$, Me , Mo , s^2 , etc... Estas medidas empíricas tienen su fundamento en las frecuencias observadas de los valores de la variable.

Después de observar el comportamiento de dicha variable durante un número elevado de meses las regularidades observadas en las frecuencias relativas permiten la definición de una distribución de probabilidad que trate de explicar el comportamiento futuro del fenómeno.

x_i	0	1	2	3	4
$p_x(x_i)$	1/3	2/5	1/6	1/15	1/30

De forma análoga al caso de la variable estadística podemos resumir los aspectos más relevantes de esta distribución mediante una serie de medidas teóricas como por ejemplo la esperanza, la mediana, la moda, la varianza, etc... Así podemos relacionar conceptos como los que se muestran en la siguiente tabla:

Medidas Empíricas	Medidas Teóricas
Frecuencia relativa	Probabilidad
Frecuencia relativa acumulada	Función de distribución
Variable estadística	Variable aleatoria
Media aritmética ($\bar{x}$)	Esperanza matemática (μ)
Varianza (s^2)	Varianza (σ^2)
Momentos no centrados (a_r)	Momentos no centrados (α_r)
Momentos centrados (m_r)	Momentos centrados (μ_r)

15.5.2. Esperanza matemática. Propiedades

EJEMPLO 15.7 Se considera el juego consistente en extraer, al azar, una carta de una baraja española de 40 cartas. Si la carta es "as" se ganan 300 unidades

monetarias, si es "figura" se ganan 100 y si se obtiene otra cualquiera se pierden 100. Se quiere calcular la ganancia media o ganancia esperada por jugada.

Sea la variable aleatoria X = "ganancia por jugada". Su correspondiente función de probabilidad es:

x_i	300	100	-100
$p_X(x_i)$	4/40	12/40	24/40

La ganancia media por jugada o ganancia esperada será

$$E[X] = 300 \frac{4}{40} + 100 \frac{12}{40} - 100 \frac{24}{40} = 0$$

- (a) Si $E[X] = 0$, entonces el juego es justo. El hecho de ganar o perder dependerá sólo del azar.
- (b) Si $E[X] < 0$, entonces sería recomendable no jugar por resultar desfavorable para el jugador.
- (c) Si $E[X] > 0$, entonces sería recomendable jugar por resultar favorable para el jugador.

Definición 15.6 Sea X una variable aleatoria discreta que toma los valores $x_1, x_2, \dots, x_n, \dots$ con probabilidades $p_X(x_i) > 0$. Llamaremos esperanza matemática, media, valor medio ó valor esperado de X , $E[X]$, a:

$$E[X] = \sum_{i=1}^{\infty} x_i p_X(x_i) = \sum_{i=1}^{\infty} x_i p[X = x_i]$$

OBSERVACIÓN 15.4 Diremos que existe la esperanza de X siempre que la serie anterior sea absolutamente convergente, es decir, cuando

$$\sum_{i=1}^{\infty} |x_i| p_X(x_i) < +\infty$$

Definición 15.7 Sea X una variable aleatoria continua con función de densidad $f_X(x)$. Se llama esperanza, media, valor medio ó valor esperado de X , $E[X]$, a:

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx.$$

OBSERVACIÓN 15.5 Diremos que existe la esperanza de X siempre que la integral anterior sea absolutamente convergente, es decir, cuando

$$\int_{-\infty}^{+\infty} |x| f_X(x) dx < +\infty.$$

OBSERVACIÓN 15.6 Es frecuente utilizar la letra griega μ para representar a $E(X)$.

Propiedades de la esperanza

Sean X e Y variables aleatorias, a y c constantes, se tienen:

- (a) $E[c] = c$.
- (b) $E[cX] = cE[X]$.
- (c) $E[X + Y] = E[X] + E[Y]$.
- (d) $E[a + cX] = a + cE[X]$.

EJEMPLO 15.8 *Se consideran las empresas de comercio al por mayor cuya producción, X , es inferior a 36000 euros. Se sabe que X , expresada en miles de euros, es una variable aleatoria con función de densidad:*

$$f_X(x) = \begin{cases} \frac{k}{\sqrt{x}}, & \text{si } x \in (0, 36) \\ 0, & \text{en otro caso} \end{cases}$$

Queremos calcular la producción esperada.

En primer lugar debemos calcular el valor de k para que $f_X(x)$ sea efectivamente función de densidad. Entonces, debe darse

$$\int_{-\infty}^{+\infty} f_X(x) dx = \int_0^{36} \frac{k}{\sqrt{x}} dx = \left[\frac{kx^{1/2}}{1/2} \right]_0^{36} = 12k = 1,$$

por tanto, $k = 1/12$.

De esta forma, la producción esperada vendrá dada por:

$$\begin{aligned} E[X] &= \int_{-\infty}^{+\infty} x f_X(x) dx = \int_0^{36} \frac{x}{12\sqrt{x}} dx = \\ &= \int_0^{36} \frac{\sqrt{x}}{12} dx = \left[\frac{x^{3/2}}{12 \cdot 3/2} \right]_0^{36} = 12 \end{aligned}$$

luego la producción esperada es de 12000 euros.

15.5.3. Momentos**Momentos no centrados de orden r**

Definición 15.8 *Sea X una variable aleatoria continua con función de densidad $f_X(x)$ ó una variable aleatoria discreta con función de probabilidad $p_X(x)$, se define el momento no centrado, respecto al origen o momento ordinario de orden*

r , α_r , como:

$$\alpha_r = E[X^r] = \begin{cases} \sum_{i=1}^{\infty} x_i^r p_X(x_i), & \text{si } X \text{ es discreta} \\ \int_{-\infty}^{+\infty} x^r f_X(x) dx, & \text{si } X \text{ es continua} \end{cases}$$

Momentos centrados de orden r

Definición 15.9 Sea X una variable aleatoria con media μ , continua con función de densidad $f_X(x)$ ó discreta con función de probabilidad $p_X(x)$. Se define su momento centrado de orden r o momento respecto de la media de orden r , μ_r , como:

$$\mu_r = E[(X - \mu)^r] = \begin{cases} \sum_{i=1}^{\infty} (x_i - \mu)^r p_X(x_i), & \text{si } X \text{ es discreta} \\ \int_{-\infty}^{+\infty} (x - \mu)^r f_X(x) dx, & \text{si } X \text{ es continua} \end{cases}$$

15.5.4. Varianza y desviación típica. Propiedades

Definición 15.10 Sea X una variable aleatoria con media μ , continua con función de densidad $f_X(x)$ ó discreta con función de probabilidad $p_X(x)$. Se llama varianza de X al momento centrado de orden 2, es decir

$$\mu_2 = \text{Var}[X] = E[(X - \mu)^2] = \begin{cases} \sum_{i=1}^{\infty} (x_i - \mu)^2 p_X(x_i), & \text{si } X \text{ es discreta} \\ \int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x) dx, & \text{si } X \text{ es continua} \end{cases}$$

OBSERVACIÓN 15.7 La varianza de la variable aleatoria X , suele representarse también por σ^2 . Se verifica que:

$$\text{Var}[X] = \sigma^2 = E[X^2] - (E[X])^2 = \alpha_2 - \alpha_1^2.$$

Definición 15.11 Sea X una variable aleatoria cualquiera, se define su desviación típica como

$$\sigma = \text{DT}[X] = +\sqrt{\text{Var}[X]}.$$

Propiedades de la varianza y de la desviación típica

Sean X variable aleatoria, a y c constantes, se verifican:

- (a) $Var[c] = 0$ ($\Rightarrow DT[c] = 0$).
- (b) $Var[cX] = c^2 Var[X]$ ($\Rightarrow DT[cX] = |c| DT[X]$).
- (c) $Var[a + X] = Var[X]$ ($\Rightarrow DT[a + X] = DT[X]$).
- (d) $Var[a + cX] = c^2 Var[X]$ ($\Rightarrow DT[a + cX] = |c| DT[X]$).

EJEMPLO 15.9 Sea X una variable aleatoria con función de densidad

$$f_X(x) = \begin{cases} \frac{-3}{4}(x^2 - 4x + 3), & \text{si } 1 < x < 3 \\ 0, & \text{en otro caso} \end{cases}$$

Queremos calcular μ y μ_2 .

(a) Calculamos la esperanza de X , μ .

$$\mu = \alpha_1 = E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx = \frac{-3}{4} \int_1^3 (x^3 - 4x^2 + 3x) dx = 2$$

(b) Calculada la esperanza, para calcular μ_2 calcularemos previamente α_2 y aplicaremos la Observación 15.7:

$$\alpha_2 = \int_{-\infty}^{+\infty} x^2 f_X(x) dx = \frac{-3}{4} \int_1^3 (x^4 - 4x^3 + 3x^2) dx = \frac{21}{5}$$

por tanto,

$$\mu_2 = \alpha_2 - \alpha_1^2 = \frac{21}{5} - 4 = \frac{1}{5}$$

EJEMPLO 15.10 Un concesionario de automóviles, A , vende 2 coches la mitad de los días y 16 la otra mitad. Otro concesionario, B , vende 8 coches la mitad de los días y 10 la otra mitad. Queremos calcular el número de coches que se espera que vendan cada uno de los concesionarios un día cualquiera y dar una medida de la representatividad de la citada medida.

Sean X_A = "número de coches que vende el concesionario A en un día " y X_B = "número de coches que vende el concesionario B en un día ".

x_i	2	16
$p_{X_A}(x_i)$	0.5	0.5

x_i	8	10
$p_{X_B}(x_i)$	0.5	0.5

$$\begin{aligned} E[X_A] &= 2 \cdot 0,5 + 16 \cdot 0,5 = 9 & E[X_B] &= 8 \cdot 0,5 + 10 \cdot 0,5 = 9 \\ Var[X_A] &= (-7)^2 \cdot 0,5 + 7^2 \cdot 0,5 = 49 & Var[X_B] &= (-1)^2 \cdot 0,5 + 1^2 \cdot 0,5 = 1 \end{aligned}$$

Ambos concesionarios venden por término medio el mismo número de coches al día, pero para el concesionario B este promedio puede considerarse más representativo ya que tiene una menor dispersión.

Capítulo 16

Modelos probabilísticos discretos

16.1. Introducción

Una de las preocupaciones de los científicos dedicados al Cálculo de Probabilidades ha sido construir modelos de distribuciones de probabilidad que pudieran representar el comportamiento teórico de diferentes fenómenos aleatorios que aparecían en el mundo real. Se puede observar como diversas distribuciones de probabilidad tienen cierta estructura parecida, es decir, responden a un modelo.

Una distribución de probabilidad queda definida mediante la especificación de la variable, su campo de variación y la determinación de sus probabilidades.

Si un conjunto de distribuciones tienen sus funciones de definición (función de distribución, de densidad, de probabilidad) con la misma estructura funcional, diremos que pertenecen a la misma familia de distribuciones o al mismo *modelo de probabilidad*.

La estructura matemática de las funciones de definición de las distribuciones de la misma familia, suele depender de uno o varios parámetros a los que llamaremos *parámetros de la distribución*.

Las ventajas de trabajar con modelos es que podemos aplicar unas fórmulas matemáticas que permiten fácilmente calcular probabilidades.

En el presente capítulo analizaremos algunos de los modelos de distribuciones de probabilidad discretos más frecuentes, de forma que podamos identificarlos cuando se presenten en un determinado experimento aleatorio.

16.2. La distribución Binomial

Consideremos un experimento aleatorio que puede dar lugar únicamente a dos resultados, A (llamado habitualmente éxito) y $\bar{A}$ (llamado habitualmente fracaso), con probabilidades de ocurrencia respectivas p y q ($p + q = 1$). Un experimento como el anterior se llama *experimento de Bernoulli*.

EJEMPLO 16.1 *Puede considerarse un experimento de Bernoulli el consistente en el lanzamiento de una moneda. En este caso los dos únicos posibles resultados serían $A = \text{"sacar cara"}$ y $\bar{A} = \text{"sacar cruz"}$ y $p = q = 1/2$.*

EJEMPLO 16.2 *Puede considerarse un experimento de Bernoulli el consistente en el lanzamiento de un dado. Podemos tomar como $A = \text{"sacar seis"}$, $\bar{A} = \text{"no sacar seis"}$, $p = 1/6$ y $q = 5/6$.*

Supongamos que se realizan n repeticiones independientes de un experimento de Bernoulli con probabilidades de éxito y fracaso respectivas p y q que permanecen invariantes a lo largo de todo el proceso.

Definición 16.1 *La variable aleatoria*

$X = \text{"número de éxitos que ocurren en las } n \text{ pruebas independientes"}$

tiene como posibles valores $X = \{0, 1, 2, \dots, n\}$ y la correspondiente función de probabilidad es

$$p[X = k] = \binom{n}{k} p^k q^{n-k}, \quad \text{para } k = 0, 1, 2, \dots, n$$

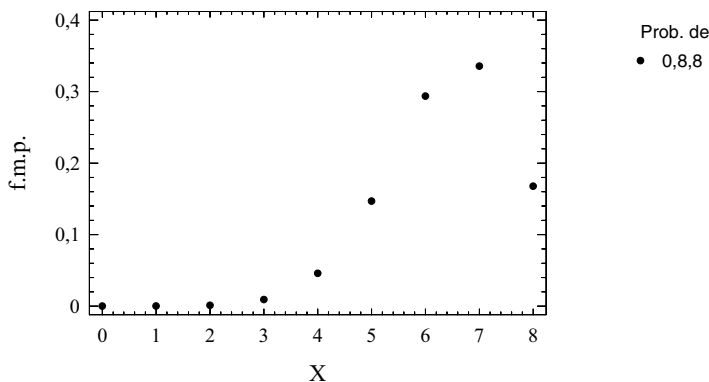
A la distribución de la variable anterior se la conoce con el nombre de distribución Binomial de parámetros n y p , que, simbólicamente representaremos por $X \rightsquigarrow \mathcal{B}(n, p)$.

Sus principales características son:

- (a) $E[X] = np$
- (b) $Var[X] = npq$
- (c) $p[X = k]$ en una distribución $\mathcal{B}(n, p)$ es igual a $p[X = n - k]$ en una distribución $\mathcal{B}(n, 1 - p)$.

La gráfica siguiente muestra la función masa de probabilidad de una distribución $\mathcal{B}(8, 0.8)$.

OBSERVACIÓN 16.1 *Una distribución Binomial de parámetro $n = 1$ se conoce también como distribución de Bernoulli de parámetro p .*



EJEMPLO 16.3 En una financiera se comprobó que el 20% de los créditos realizados resultaban impagados. Si se seleccionan al azar ocho créditos, queremos calcular:

(a) La probabilidad de que exactamente cinco créditos sean pagados.

Sea $X =$ " número de créditos pagados de los ocho seleccionados "

$$X \rightsquigarrow B(8, 0,8)$$

$$p[X = 5] = \binom{8}{5} (0,8)^5 (0,2)^3 = 0,147$$

(b) La probabilidad de que se paguen al menos tres.

$$\begin{aligned} p[X \geq 3] &= 1 - p[X < 3] = 1 - (p[X = 0] + p[X = 1] + p[X = 2]) = \\ &= 1 - \left[\binom{8}{0} (0,8)^0 (0,2)^8 + \binom{8}{1} (0,8)^1 (0,2)^7 + \binom{8}{2} (0,8)^2 (0,2)^6 \right] = 0,998 \end{aligned}$$

(c) Número de créditos, de esos ocho, que se esperan cobrar.

$$E[X] = np = 8(0,8) = 6,4$$

En un grupo de ocho créditos se cobran 6,4 de promedio.

Definición 16.2 Se dice que una variable verifica el teorema de adición para algunos de sus parámetros o que es reproductiva respecto de ellos, si dadas dos o más variables aleatorias independientes que siguen todas ellas un mismo modelo de distribución, con parámetros iguales o distintos, la variable suma de todas sigue también una distribución del mismo tipo, con parámetros iguales a la suma de los parámetros respecto de los que sea reproductiva.

Teorema 16.1 Sean $X \rightsquigarrow \mathcal{B}(n_1, p)$ e $Y \rightsquigarrow \mathcal{B}(n_2, p)$ independientes entre si. Se verifica que

$$X + Y \rightsquigarrow \mathcal{B}(n_1 + n_2, p)$$

Es decir, la distribución Binomial verifica el teorema de adición o es reproductiva respecto del parámetro n .

EJEMPLO 16.4 Consideremos el experimento aleatorio consistente en el lanzamiento de una moneda no cargada. Supongamos que repetimos dicho lanzamiento tres veces. A su vez, lanzamos dos veces, y de forma independiente, otra moneda no cargada distinta. Estamos interesados en estudiar la probabilidad de obtener al menos dos caras entre el total de los cinco lanzamientos realizados con ambas monedas. Entonces, podemos definir las variables aleatorias

$$X = \text{"n}^0 \text{ de caras obtenidas en tres lanzamientos de la moneda-1"} \rightsquigarrow \mathcal{B}(3, 0,5)$$

$$Y = \text{"n}^0 \text{ de caras obtenidas en dos lanzamientos de la moneda-2"} \rightsquigarrow \mathcal{B}(2, 0,5)$$

siendo éxito, en este caso, el suceso $A = \text{"obtener cara"}'$, con $p(A) = 0,5$.

Por tanto,

$$W = X + Y \rightsquigarrow \mathcal{B}(5, 0,5)$$

y la probabilidad que queremos calcular es

$$\begin{aligned} p[W \geq 2] &= 1 - (p[W = 0] + p[W = 1]) = \\ &= 1 - \binom{5}{0}(0,5)^0(0,5)^5 - \binom{5}{1}(0,5)^1(0,5)^4 = 0,8125 \end{aligned}$$

16.3. La distribución de Poisson

Supongamos que se realiza un experimento consistente en observar la aparición de ciertos acontecimientos puntuales o éxitos que ocurren sobre un soporte continuo (tiempo, espacio, longitud, etc...) con las siguientes condiciones:

- (a) El número medio de éxitos a largo plazo es constante.
- (b) Los éxitos ocurren aleatoriamente de forma independiente.

A este tipo de experimentos se les llama *procesos de Poisson* y son ejemplos del mismo la llegada de clientes a cierta ventanilla de un banco en una hora, los defectos que aparecen en cada rollo de cable, etc.

Definición 16.3 Para este tipo de procesos, podemos definir la variable

$$X = \text{"número de éxitos en un intervalo de amplitud determinada"}$$

que puede tomar como posibles valores $X = \{0, 1, 2, \dots\}$ con función de probabilidad

$$p[X = k] = e^{-\lambda} \cdot \frac{\lambda^k}{k!}, \quad \text{para } k = 0, 1, 2, \dots$$

Diremos que una variable de este tipo sigue una distribución de Poisson de parámetro λ ($\lambda > 0$) y escribiremos $X \rightsquigarrow \mathcal{P}(\lambda)$.

OBSERVACIÓN 16.2

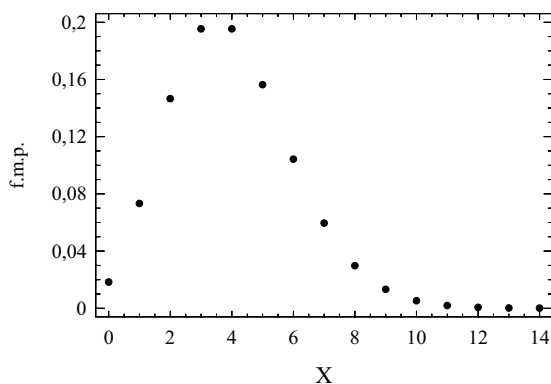
(a) En 1898 Bortkiewicz demostró que el número de soldados muertos por coces de mula en un año respondía al modelo de una variable de Poisson, donde en los 6 primeros valores se encuentra el 95.5 % de la probabilidad. Si bien dicho hecho era probable en un número reducido de ocasiones era muy improbable que fuera frecuente.

(b) La distribución de Poisson se llama "Ley de los Sucesos Raros" ya que trabaja con sucesos éxito con probabilidades muy cercanas a cero.

Sus principales características son:

(a) $E[X] = \lambda$

(b) $Var[X] = \lambda$



La anterior gráfica muestra la función masa de probabilidad de una distribución $\mathcal{P}(4)$.

EJEMPLO 16.5 Se ha observado que el número medio de clientes que entran en cierto comercio en una hora es de cuatro. Estamos interesados en calcular:

(a) El porcentaje de las horas en las que no llega nadie.

Sea la variable

$X =$ "nº de clientes que entran en el comercio en una hora" $\rightsquigarrow \mathcal{P}(4)$

$$p[X = 0] = e^{-4} \cdot \frac{4^0}{0!} = e^{-4} = 0,0183$$

El 1,83 % de las horas, no entra nadie en el comercio.

(b) La probabilidad de que en una hora lleguen entre tres y seis clientes.

$$\begin{aligned} p[3 \leq X \leq 6] &= p[X = 3] + p[X = 4] + p[X = 5] + p[X = 6] = \\ &= e^{-4} \cdot \frac{4^3}{3!} + e^{-4} \cdot \frac{4^4}{4!} + e^{-4} \cdot \frac{4^5}{5!} + e^{-4} \cdot \frac{4^6}{6!} = 0,6513 \end{aligned}$$

Teorema 16.2 Sean $X \rightsquigarrow \mathcal{P}(\lambda_1)$ e $Y \rightsquigarrow \mathcal{P}(\lambda_2)$ independientes entre si. Se verifica que

$$X + Y \rightsquigarrow \mathcal{P}(\lambda_1 + \lambda_2)$$

Es decir, la distribución de Poisson verifica el teorema de adición o es reproductiva respecto del parámetro λ .

EJEMPLO 16.6 Supongamos que el comercio del Ejemplo 16.5 permanece abierto cuatro horas cada día. Queremos calcular la probabilidad de que, en un día, lleguen diez clientes.

Sea la variable

$X_i =$ "nº de clientes que entran en el comercio en la i -ésima hora" $\rightsquigarrow \mathcal{P}(4)$

Por tanto, el número total de clientes que llegan al comercio en un día viene dado por la variable

$$Y = \sum_{i=1}^4 X_i \rightsquigarrow \mathcal{P}(16)$$

Entonces

$$p[Y = 10] = e^{-16} \cdot \frac{16^{10}}{10!} = 0,0341$$

16.3.1. Aproximación de Poisson a la distribución Binomial

Inicialmente Poisson determinó la distribución que lleva su nombre como límite de la Binomial de la siguiente forma, es decir, si el número de elementos observados es muy grande y la probabilidad de observar la característica estudiada en cada elemento es muy pequeña, pero de manera que el número medio de sucesos permanezca constante:

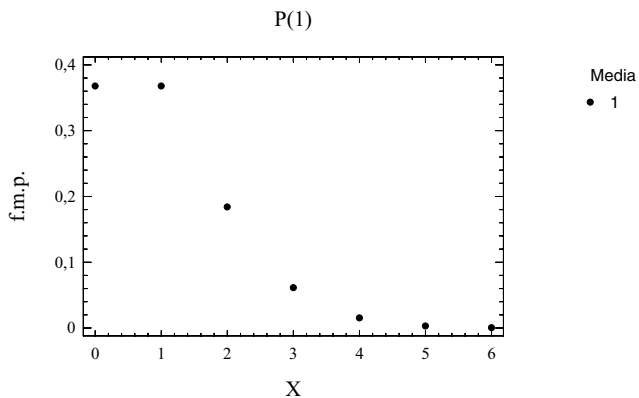
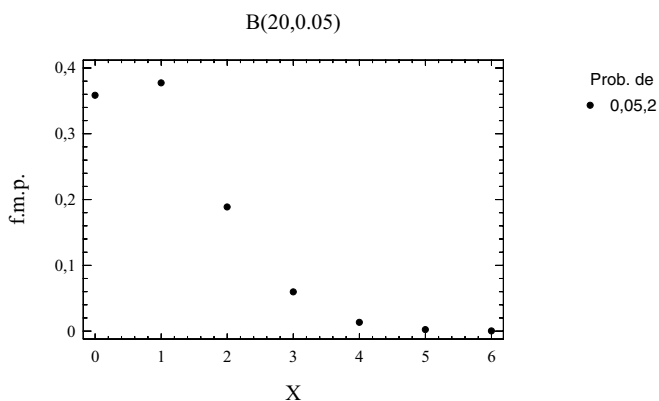
$$\lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0 \\ np \rightarrow \lambda > 0}} \binom{n}{k} p^k (1-p)^{n-k} = e^{-np} \cdot \frac{(np)^k}{k!} = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

Este resultado podemos concretarlo en el siguiente:

Teorema 16.3 Sea $X \rightsquigarrow \mathcal{B}(n, p)$. Se verifica que si $n \rightarrow \infty$, $p \rightarrow 0$, $np \rightarrow \lambda > 0$, la distribución de X tiende a ser $\mathcal{P}(\lambda)$.

OBSERVACIÓN 16.3 En la práctica se obtiene una aproximación aceptable si

$$p \leq 0,1 \text{ y } np = \lambda < 5$$



EJEMPLO 16.7 *Se sabe que la probabilidad de que una determinada máquina fabrique una pieza defectuosa es 0,0001. Si en un año se fabrican 20000 piezas, ¿cuál es la probabilidad de que el número de piezas defectuosas producidas en un año sea mayor que dos?*

Sea X = "número de piezas defectuosas entre 20000 ", entonces

$$X \rightsquigarrow \mathcal{B}(20000, 0,0001)$$

Puesto que $p = 0,0001 \leq 0,1$ y $np = 2 < 5$, podemos aproximar la ley Binomial por una Poisson, luego

$$X \rightsquigarrow \mathcal{P}(2) \text{ (aproximadamente)}$$

$$\begin{aligned} p[X > 2] &= 1 - p[X \leq 2] = 1 - (p[X = 0] + p[X = 1] + p[X = 2]) \approx \\ &\approx 1 - 0,1353 - 0,2707 - 0,2707 = 1 - 0,6767 = 0,3233 \end{aligned}$$

Capítulo 17

La distribución Normal

17.1. Introducción

Esta distribución fue considerada por primera vez por Abraham De Moivre, en 1753, en relación con la discusión y estudio de formas límites de otras distribuciones. Pronto pasó al olvido y no fue "redescubierta" hasta principios del siglo XIX por Gauss y Laplace en sus trabajos sobre la teoría de errores de observación.

El nombre de "Normal" posee sólo un carácter histórico ya que se creyó, más o menos axiomáticamente durante mucho tiempo, que todas las distribuciones eran de este tipo o se acercaban a ella como forma límite ideal.

Las demás distribuciones eran, en consecuencia, "anormales". Aunque este punto de vista fue, sin duda, exagerado y sufrió posteriormente modificaciones considerables, queda fuera de toda duda que en un gran número de aplicaciones importantes encontramos la distribución Normal o aproximadamente Normal.

Podemos resumir la importancia de la distribución Normal diciendo que:

- (a) Un gran número de fenómenos reales pueden modelizarse con ella. Por ejemplo, las medidas físicas del cuerpo humano en una población, las características psíquicas medidas por tests de inteligencia o personalidad, las medidas de calidad en muchos procesos industriales o los errores de las observaciones astronómicas.
- (b) Todas aquellas variables que puedan considerarse causadas por un gran número de pequeños efectos tienden a distribuirse como una distribución Normal.
- (c) Muchas otras distribuciones pueden aproximarse mediante la distribución Normal.
- (d) Es de gran interés por la simplicidad de sus propiedades y características.

17.2. Definición y propiedades

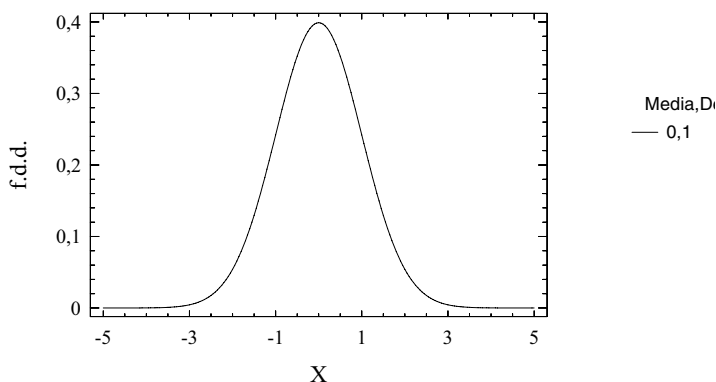
Definición 17.1 Se dirá que la variable aleatoria X sigue una distribución Normal de parámetros μ y σ si su función de densidad es de la forma:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \quad x \in \mathbb{R}, \sigma > 0, \mu \in \mathbb{R}$$

Simbólicamente escribiremos $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$.

La representación gráfica de su función de densidad presenta los siguientes aspectos de interés:

- (a) Es estrictamente positiva.
- (b) Máximo: $\left(\mu, \frac{1}{\sigma\sqrt{2\pi}}\right)$.
- (c) Puntos de inflexión: $\left(\mu \pm \sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$.
- (d) Es simétrica respecto de la recta $x = \mu$.
- (e) Es creciente si $x < \mu$ y decreciente si $x > \mu$.
- (f) Tiene a la recta $y = 0$ como asíntota horizontal.

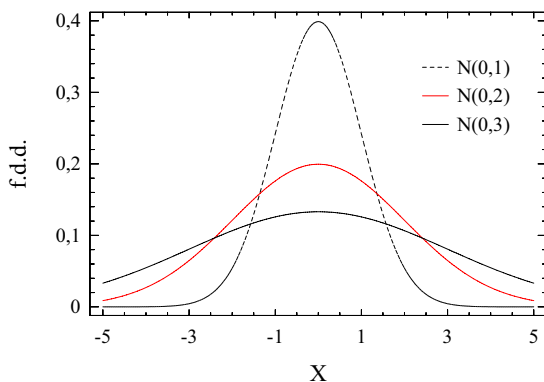
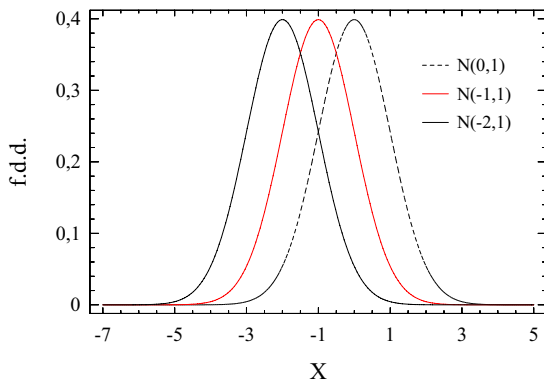


La anterior gráfica muestra la función de densidad de una distribución $\mathcal{N}(0, 1)$. Las principales características de una distribución $\mathcal{N}(\mu, \sigma)$ son:

(a) $E[X] = M_o = M_e = \mu$

(b) $Var[X] = \sigma^2$

En las representaciones gráficas siguientes se muestran las funciones de densidad de distintas distribuciones normales y en ellas observamos el efecto que los parámetros μ y σ tienen sobre las mismas.



Teorema 17.1 Sean $X \rightsquigarrow \mathcal{N}(\mu_1, \sigma_1)$ e $Y \rightsquigarrow \mathcal{N}(\mu_2, \sigma_2)$ independientes entre sí, se verifica que

$$X + Y \rightsquigarrow \mathcal{N}\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

Es decir, la distribución Normal verifica el teorema de adición o es reproductiva respecto de los parámetros μ y σ^2 .

EJEMPLO 17.1 *Un estudio de arquitectura tiene tres tipos de ingresos anuales independientes: por tarifa de edificación, por tarifa de urbanismo y por peritaciones. Se sabe que cada uno de estos ingresos, medidos en miles de euros, se distribuye normalmente, con distribuciones respectivas $\mathcal{N}(84, 12)$, $\mathcal{N}(72, 24)$ y $\mathcal{N}(18, 12)$. Queremos calcular el porcentaje de los años en los que los ingresos totales por edificación y por urbanismo no superan los 150000 euros.*

Consideremos las variables

$X =$ "ingresos anuales, en miles de euros, por tarifa de edificación"

$$X \rightsquigarrow \mathcal{N}(84, 12)$$

$Y =$ "ingresos anuales, en miles de euros, por tarifa de urbanismo" $\rightsquigarrow \mathcal{N}(72, 24)$,

independientes entre sí, entonces

$$X + Y \rightsquigarrow \mathcal{N}(156, \sqrt{12^2 + 24^2}) \equiv \mathcal{N}(156, 6\sqrt{20})$$

y la probabilidad que necesitaríamos calcular sería $p[X + Y \leq 150]$.

Teorema 17.2 *Sean $X \rightsquigarrow \mathcal{N}(\mu_1, \sigma_1)$ e $Y \rightsquigarrow \mathcal{N}(\mu_2, \sigma_2)$ independientes entre sí, se verifica que*

$$aX + bY + c \rightsquigarrow \mathcal{N}\left(a\mu_1 + b\mu_2 + c, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}\right); \quad a, b, c \in \mathbb{R}$$

EJEMPLO 17.2 *Considerando el enunciado del Ejemplo 17.1 queremos calcular el porcentaje de los años en los que los ingresos anuales por tarifa de urbanismo superan a los ingresos anuales por peritaciones en más de 52000 euros.*

Sean

$Y =$ "ingresos anuales, en miles de euros, por tarifa de urbanismo" $\rightsquigarrow \mathcal{N}(72, 24)$,

$W =$ "ingresos anuales, en miles de euros, por peritaciones" $\rightsquigarrow \mathcal{N}(18, 12)$,

independientes entre sí, entonces

$$Y - W \rightsquigarrow \mathcal{N}(54, \sqrt{24^2 + 12^2}) \equiv \mathcal{N}(54, 6\sqrt{20})$$

y la probabilidad que necesitaríamos calcular sería

$$p[Y > W + 52] = p[Y - W > 52]$$

17.3. Distribución Normal tipificada

Para calcular la probabilidad de que una variable aleatoria $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$ se encuentre en un determinado intervalo se usan los valores tabulados correspondientes a la variable

$$Z = \frac{X - \mu}{\sigma} \rightsquigarrow \mathcal{N}(0, 1)$$

que recibe el nombre de distribución Normal Tipificada.

Entre otras propiedades destacamos las siguientes:

- (a) No depende de ningún parámetro.
- (b) $E[Z] = 0$.
- (c) $Var[Z] = 1$.
- (d) Su función de densidad es simétrica respecto del eje $z = 0$, tiene su máximo en $z = 0$ y dos puntos de inflexión en $z = 1$ y $z = -1$.
- (e) La distribución Normal tipificada, Z , es la única distribución Normal que está tabulada.

17.4. Uso de tablas

Definición 17.2 Representaremos por z_α al valor de la abscisa que deja a su derecha un área bajo la curva normal igual a α (nivel de significación). Este valor se llama también punto crítico y es el que verifica que

$$p[Z \geq z_\alpha] = \alpha$$

EJEMPLO 17.3 Un profesor realiza un test de 100 preguntas a un curso de 200 alumnos. Suponiendo que las puntuaciones, X , obtenidas por los alumnos siguen una distribución Normal de media 60 puntos y desviación típica 10 puntos, se quieren calcular las siguientes probabilidades:

- | | | |
|----------------------------|----------------------------|------------------------------|
| (a) $p[X \geq 70]$ | (b) $p[X \leq 80]$ | (c) $p[X \leq 30]$ |
| (d) $p[X \geq 46]$ | (e) $p[39 \leq X \leq 80]$ | (f) $p[80 \leq X \leq 82,5]$ |
| (g) $p[30 \leq X \leq 40]$ | | |

Consideramos la variable X "puntuación obtenida" $\rightsquigarrow \mathcal{N}(60, 10)$.
La correspondiente variable tipificada es:

$$Z = \frac{X - 60}{10} \rightsquigarrow \mathcal{N}(0, 1)$$

Entonces

(a)

$$p[X \geq 70] = p\left[\frac{X - 60}{10} \geq \frac{70 - 60}{10}\right] = p[Z \geq 1] = 0,1587$$

(b)

$$\begin{aligned} p[X \leq 80] &= p\left[\frac{X - 60}{10} \leq \frac{80 - 60}{10}\right] = p[Z \leq 2] = \\ &= 1 - p[Z > 2] = 1 - 0,0228 = 0,9772 \end{aligned}$$

(c)

$$p[X \leq 30] = p\left[\frac{X - 60}{10} \leq \frac{30 - 60}{10}\right] = p[Z \leq -3] = P[Z \geq 3] = 0,00135$$

(d)

$$\begin{aligned} p[X \geq 46] &= p\left[\frac{X - 60}{10} \geq \frac{46 - 60}{10}\right] = p[Z \geq -1,4] = \\ &= p[Z \leq 1,4] = 1 - p[Z > 1,4] = 1 - 0,0808 = 0,9192 \end{aligned}$$

(e)

$$\begin{aligned} p[39 \leq X \leq 80] &= p\left[\frac{39 - 60}{10} \leq \frac{X - 60}{10} \leq \frac{80 - 60}{10}\right] = p[-2,1 \leq Z \leq 2] = \\ &= 1 - (p[Z > 2] + p[Z < -2,1]) = 1 - p[Z > 2] - p[Z > 2,1] = \\ &= 1 - 0,0228 - 0,0179 = 0,9593 \end{aligned}$$

(f)

$$\begin{aligned} p[80 \leq X \leq 82,5] &= p\left[\frac{80 - 60}{10} \leq \frac{X - 60}{10} \leq \frac{82,5 - 60}{10}\right] = \\ &= p[2 \leq Z \leq 2,25] = p[Z \geq 2] - p[Z > 2,25] = 0,0228 - 0,0122 = 0,0106 \end{aligned}$$

(g)

$$\begin{aligned} p[30 \leq X \leq 40] &= p\left[\frac{30 - 60}{10} \leq \frac{X - 60}{10} \leq \frac{40 - 60}{10}\right] = p[-3 \leq Z \leq -2] = \\ &= p[2 \leq Z \leq 3] = p[Z \geq 2] - p[Z > 3] = 0,0228 - 0,00135 = 0,02145 \end{aligned}$$

EJEMPLO 17.4 Consideremos las probabilidades que se querían obtener en los Ejemplos 17.1 y 17.2:

- (a) $p[X + Y \leq 150] = p[Z \leq -0,2236] = p[Z \geq 0,2236] = 0,4129$, teniendo en cuenta que $X + Y \rightsquigarrow \mathcal{N}(156, 6\sqrt{20})$. Por tanto, el 41,29% de los años, el total de los ingresos anuales por edificación y urbanismo no superan los 150000 euros.

- (b) $p[Y - W > 52] = p[Z > -0,0745] = 1 - p[Z \geq 0,0745] = 1 - 0,4721 = 0,5279$, siendo $Y - W \rightsquigarrow \mathcal{N}(54, 6\sqrt{20})$. Por tanto, el 52,79 % de los años los ingresos anuales por tarifa de urbanismo superan a los ingresos por peritaciones en más de 52000 euros.

17.5. Teorema Central del Límite

El nombre de Teorema Central del Límite fue usado por primera vez por George Pólya, en un artículo publicado en 1920, con el objeto de insistir en que éste era el problema central a que se dedicaban los investigadores desde el siglo XVIII.

Teorema 17.3 Sea $X_1, X_2, X_3, \dots$ una sucesión de variables aleatorias independientes, idénticamente distribuidas y con momentos de segundo orden finitos. Si se tiene que $E[X_i] = \mu$ y $\text{Var}[X_i] = \sigma^2$, $\forall i$, se verifica que:

$$\sum_{i=1}^n X_i \rightsquigarrow \mathcal{N}(n\mu, \sigma\sqrt{n})$$

cuando n tiende a ∞ .

OBSERVACIÓN 17.1 En la práctica se obtiene una aproximación buena si $n > 30$.

Viene a expresar el hecho de que, bajo condiciones bastante generales, la suma de variables aleatorias independientes idénticamente distribuidas tiende a distribuirse según una distribución Normal. La primera demostración rigurosa la dio, en 1901, el matemático ruso Alexander M. Liapunov.

EJEMPLO 17.5 En un proceso de fabricación se sabe que el número de unidades defectuosas producidas diariamente se distribuye según un distribución de Poisson de media 10 unidades. Si el número de unidades defectuosas producidas en un día es independiente de las producidas en los restantes días, queremos calcular la probabilidad de que en 150 días, se produzcan entre 1480 y 1510 unidades defectuosas.

Consideremos las variables $X_1, X_2, \dots, X_{150}$, siendo

$X_i = \text{"número de unidades defectuosas producidas el día } i\text{-ésimo"} \rightsquigarrow \mathcal{P}(10)$

La variable

$$Y = \sum_{i=1}^{150} X_i = \text{"número de unidades defectuosas producidas en 150 días"}$$

$$Y \rightsquigarrow \mathcal{N}(1500, \sqrt{1500})$$

Ello se produce aproximadamente en virtud del Teorema Central del Límite. Por tanto, la probabilidad que deseamos calcular será:

$$\begin{aligned} p[1480 \leq Y \leq 1510] &\approx p[-0,5163 \leq Z \leq 0,2581] = \\ &= 1 - p[Z > 0,5163] - p[Z > 0,2581] = 1 - 0,3015 - 0,3974 = 0,3011 \end{aligned}$$

17.6. Aproximaciones mediante la Normal

En el capítulo anterior veíamos como, bajo ciertas condiciones, la distribución binomial podía ser aproximada por la distribución de Poisson. En el presente capítulo estudiaremos dos aproximaciones a la distribución Normal.

(a) Aproximación de la distribución Binomial mediante la distribución Normal.

Teorema 17.4 (De Moivre – Laplace) *Sea X una variable aleatoria con distribución $\mathcal{B}(n, p)$. Se verifica que si $n \rightarrow \infty$, la distribución de X tiende a ser $\mathcal{N}(np, \sqrt{npq})$.*

OBSERVACIÓN 17.2 *En la práctica se obtiene una aproximación aceptable si*

$$p < 0,1 \quad y \quad np > 5 \quad \text{ó} \quad 0,1 < p < 0,9 \quad y \quad n > 30$$

EJEMPLO 17.6 *Un agente de bolsa comprobó que la probabilidad de que una determinada acción subiera de valor después de una semana era 0,5. Si el citado agente compró 100 acciones. Queremos calcular:*

(i) *La probabilidad de que suban más de 65 acciones.*

Sea X = "número de acciones, entre las 100 compradas, que suben"

$$X \rightsquigarrow \mathcal{B}(100, 0,5)$$

Observemos que se verifican las condiciones del Teorema de De Moivre-Laplace

$$0,1 < p = 0,5 < 0,9 \quad y \quad n = 100 > 30$$

por tanto,

$$X \rightsquigarrow \mathcal{N}\left(100(0,5), \sqrt{100(0,5)(0,5)}\right) \equiv \mathcal{N}(50, 5) \quad (\text{aproximadamente})$$

Tras tipificar obtenemos

$$Z = \frac{X - 50}{5} \rightsquigarrow \mathcal{N}(0, 1)$$

y la probabilidad que queremos calcular es

$$p[X > 65] \approx p\left[\frac{X - 50}{5} > \frac{65 - 50}{5}\right] = p[Z > 3] = 0,00135$$

(ii) La probabilidad de que suban más de 40 y menos de 60 acciones.

$$p[40 < X < 60] \approx p\left[\frac{40-50}{5} < \frac{X-50}{5} < \frac{60-50}{5}\right] =$$

$$= p[-2 < Z < 2] = 1 - 2p[Z \geq 2] = 1 - 2(0,0228) = 0,9544$$

(iii) La probabilidad de que suban menos de 30 acciones.

$$p[X < 30] \approx p[Z < -4] = p[Z > 4] = 0,0000317$$

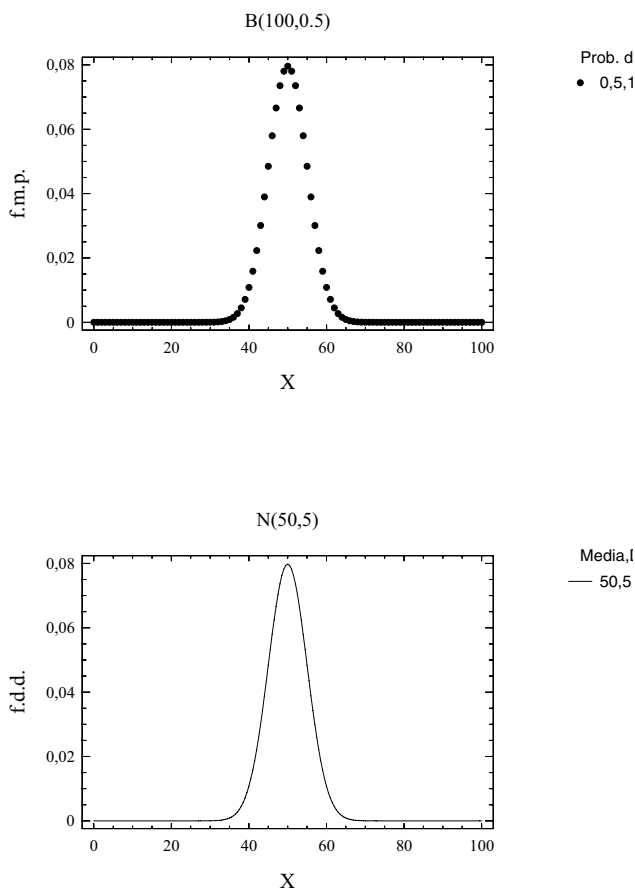


Figura 17.1: Comparación entre las funciones de masa de probabilidad y de densidad de las distribuciones $\mathcal{B}(100, 0.5)$ y $\mathcal{N}(50, 0.5)$.

- (b) Aproximación de la distribución de Poisson mediante la distribución Normal.

Teorema 17.5 Sea X una variable aleatoria con distribución $\mathcal{P}(\lambda)$. Se verifica que si $\lambda \rightarrow \infty$, la distribución de X tiende a ser $\mathcal{N}(\lambda, \sqrt{\lambda})$.

OBSERVACIÓN 17.3 En la práctica se obtiene una aproximación aceptable si $\lambda > 10$.

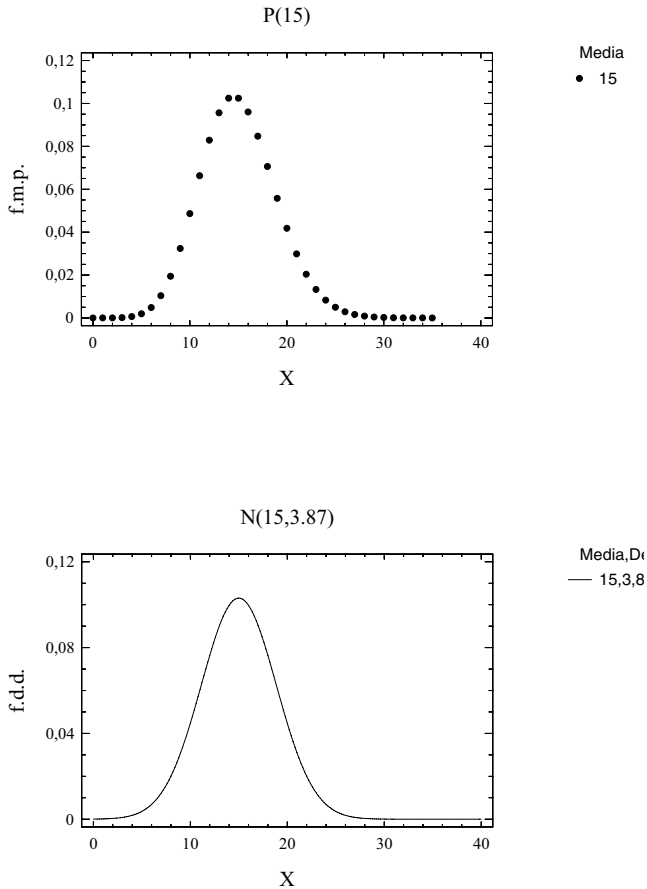


Figura 17.2: Comparación entre las funciones de masa de probabilidad y de densidad de las distribuciones $\mathcal{P}(15)$ y $\mathcal{N}(15, 3.87)$.

EJEMPLO 17.7 *Un servicio técnico de reparaciones ha observado que recibe, por término medio, 15 llamadas diarias. Queremos obtener la probabilidad de que se reciban más de 20 llamadas en un día.*

Sea X = "número de llamadas recibidas en un día" $\rightsquigarrow \mathcal{P}(15)$

Como $\lambda = 15 > 10$, aplicando el Teorema 17.5 se obtiene

$$X \rightsquigarrow \mathcal{N}\left(15, \sqrt{15}\right) \equiv \mathcal{N}(15, \sqrt{15}) \text{ (aproximadamente)}$$

y la probabilidad que queremos calcular

$$p[X > 20] \approx p[Z > 1,2919] = 0,0985$$

Cuestiones, problemas y recursos

Cuestiones

1. Representa gráficamente una función que pudiera ser función de distribución de una variable aleatoria discreta, comentando sobre ella las principales propiedades conocidas de toda función de distribución.
2. Representa gráficamente una función que pudiera ser función de distribución de una variable aleatoria continua, comentando sobre ella las principales propiedades conocidas de toda función de distribución.
3. A continuación se proporciona la función de distribución de cierta variable aleatoria:

$$F_x(x) = \begin{cases} 0, & \text{si } x < -2 \\ 0,3, & \text{si } -2 \leq x < 2 \\ 0,8, & \text{si } 2 \leq x < 4 \\ 1, & \text{si } 4 \leq x \end{cases}$$

- (a) Representala gráficamente y comenta a qué tipo de variable pertenece y descríbela.
 - (b) Calcula y representa la función masa de probabilidad.
4. A la vista de la representación gráfica de la función F_x , razona si puede ser la función de distribución de cierta variable aleatoria (enumera todos los argumentos en favor o en contra)

$$F_x(x) = \begin{cases} 0, & \text{si } x < 0 \\ 3 - 0,3x, & \text{si } 0 \leq x < 10 \\ 1, & \text{si } 10 \leq x \end{cases}$$

5. Razona si la siguiente función puede ser función de densidad de una variable

aleatoria X ,

$$f_X(x) = \begin{cases} 2x - 10, & \text{si } 4 < x < 5 \\ x - 5, & \text{si } 5 \leq x < 7. \\ 0, & \text{en el resto} \end{cases}$$

6. Comenta con todo detalle si puede ser función de densidad de una variable aleatoria la función:

$$f_X(x) = \begin{cases} \frac{x}{40}, & \text{si } 0 < x < 10 \\ 0, & \text{en el resto.} \end{cases}$$

7. La demanda de cierto artículo presenta la siguiente función de masa de probabilidad:

x_i	0	1	2	3	4	5	6
$p_X(x_i)$	p	0,2	0,2	0,24	0,2	p	q

Sabiendo que la demanda media es 2,73, calcula los valores de p y q .

8. La variable aleatoria X toma tres valores según la siguiente función de masa de probabilidad: $p[X = 1] = p$, $p[X = 3] = 0,2$, $p[X = 5] = p$. Calcula la varianza de la variable X .
9. ¿Cuántas quejas se pueden esperar por día en la oficina de una empresa, si las probabilidades de que reciba 0, 1, 2, 3, 4, 5, 6, 7 u 8 quejas al día son de 0.06, 0.21, 0.24, 0.18, 0.14, 0.10, 0.04, 0.02 y 0.01, respectivamente?
10. La probabilidad de que un inversionista pueda vender una acción con beneficio de 2500 euros, un beneficio de 1500 euros, un beneficio de 500 euros o una pérdida de 500 euros son 0.22, 0.36, p y q , respectivamente. Calcula p y q para que el beneficio esperado sea de 1160 euros.
11. A un individuo se le propone el siguiente juego: de una baraja de 40 cartas se extrae una carta al azar. Si la carta es as gana 800 euros, si la carta es figura gana 100 euros y si obtiene otra cualquiera pierde 100 euros. ¿Le recomendaría jugar en el juego?
12. Se plantea el juego de azar consistente en lanzar un dado no cargado. Si sale 1 el jugador gana 30 céntimos de euro. Si sale 2, 3, 4 ó 5, el jugador pierde 20 céntimos de euro. Si sale 6 el jugador gana 50 céntimos de euro. Calcula la ganancia esperada del juego propuesto. Interpreta el resultado obtenido.
13. ¿Por qué a la distribución de Poisson se le llama "Ley de los Sucesos Raros"?

14. Razona sin mirar en las tablas y sin realizar ningún cálculo cuál es el valor de las siguientes probabilidades: (a) $p(X < 0)$, si $X \rightsquigarrow \mathcal{N}(0, 0,3)$; (b) $p(Y \leq 5)$, si $Y \rightsquigarrow \mathcal{B}(5, 0,3)$; (c) $p(Y \leq -1)$, si $Y \rightsquigarrow \mathcal{P}(3)$.
15. ¿Para qué distribuciones de las siguientes, y por qué, es cierta la igualdad $p(X \geq 2) = p(X > 2)$?: (a) $X \rightsquigarrow \mathcal{N}(0, 1)$; (b) $X \rightsquigarrow \mathcal{B}(5, 0,1)$.
16. Sean X e Y dos variables aleatorias independientes que se distribuyen según $\mathcal{P}(30)$ y $\mathcal{B}(50, 0,3)$, respectivamente. Razona qué distribución sigue, aproximadamente, la variable $X + Y$.
17. Sean X e Y dos variables aleatorias independientes que se distribuyen según $\mathcal{N}(4, 1)$ y $\mathcal{B}(50, 0,3)$, respectivamente. Razona qué distribución sigue, aproximadamente, la variable $X + Y$.
18. Sean X e Y dos variables aleatorias independientes que se distribuyen según $\mathcal{N}(1, 2)$ y $\mathcal{N}(-4, 1)$, respectivamente. Razona qué distribución sigue cada una de las siguientes variables: (a) $3X + 1$; (b) $X - Y$.
19. Sean $X_1, X_2, \dots, X_{49}$ variables aleatorias independientes e idénticamente distribuidas con $E[X_i] = 0$ y $Var[X_i] = 25$, $i = 1, 2, \dots, 49$. Explica razonadamente la distribución aproximada que sigue $\sum_{i=1}^{49} X_i$.
20. El 30 % de los artículos fabricados por cierta máquina son defectuosos. Si se eligen 10 artículos al azar, ¿cuál es la probabilidad de que al menos tres sean defectuosos?
21. En la centralita telefónica de un pueblo se reciben un promedio de 12 llamadas por hora. Queremos conocer la probabilidad de que en un intervalo de 10 minutos se produzcan de 3 a 5 llamadas. Plantea qué distribución es conveniente para obtener dicha probabilidad (Define la variable, indica su distribución y plantea la probabilidad que se quiere conocer)
22. La proporción de individuos de una población con renta superior a doce mil euros es de 0.4 %. Se quiere determinar la probabilidad de que, entre 600 individuos consultados haya exactamente 2 con ese nivel de renta (Define la variable adecuadamente, indica su distribución y plantea la probabilidad que se quiere conocer)
23. Enuncia el Teorema Central del Límite.
24. Cita las causas que hacen que la distribución Normal sea tan importante.

Problemas

1. La demanda diaria, X , de un cierto artículo es una variable aleatoria con la siguiente función de masa de probabilidad:

x_i	1	2	3	4
$p_X(x_i)$	$2p$	$2p$	$\frac{4p}{3}$	$\frac{2p}{3}$

- Comprueba que $p = 1/6$ es el valor que hace que $p_X(x)$ sea función de masa de probabilidad.
 - Calcula y representa la función de distribución.
 - ¿Cuál es la demanda diaria esperada? ¿Y la desviación típica de la demanda diaria?
2. El número de coches que vende en una semana un empleado comercial de cierto concesionario de automóviles de lujo, X , así como sus respectivas probabilidades aparecen en la siguiente tabla:

x_i	0	1	2	3	4	5
$p_X(x_i)$	$2p$	$6p$	$6p$	$4p$	p	p

- Calcula el valor de p que hace que $p_X(x)$ sea la función de masa de probabilidad de la variable aleatoria X .
 - Obtégase la función de distribución de la variable X en el intervalo $[2, 5)$.
 - Calcula $p(2 < X \leq 4)$.
 - Calcula el número medio de coches que vende semanalmente un empleado y la varianza.
3. Sea una variable aleatoria discreta con función de probabilidad dada por la siguiente tabla:

x_i	2	4	6	8	10	12
$p_X(x)$	p	$5p$	$3p$	$6p$	$3p$	$2p$

Se pide:

- Determina el valor de p para que el conjunto de probabilidades dado sea función de probabilidad de la variable X .
- Calcula y representa la función de distribución de la variable.
- Calcula la esperanza de X . ¿Qué porcentaje de valores superan este valor?

4. Sea la variable aleatoria X definida por su función de densidad

$$f_X(x) = \begin{cases} k, & \text{si } 0 \leq x < 1 \\ kx, & \text{si } 1 \leq x < 2 \\ 0, & \text{en el resto} \end{cases}$$

- (a) Calcula k para que $f_X(x)$ sea función de densidad.
 - (b) Calcula y representa la función de distribución.
5. Sea la variable aleatoria continua X definida por su función de densidad

$$f_X(x) = \begin{cases} 1/6, & \text{si } 1 \leq x \leq 2 \\ k, & \text{si } 2 < x \leq 3 \\ 0, & \text{en el resto} \end{cases}$$

- (a) Calcula k para que $f_X(x)$ sea función de densidad.
 - (b) Calcula la función de distribución, (suponga $k=5/6$).
 - (c) Calcula la media.
 - (d) Calcula $p(1,75 \leq X \leq 2,5)$.
6. Sea la variable aleatoria X definida por su función de densidad

$$f_X(x) = \begin{cases} x + k, & \text{si } 0 \leq x \leq 1 \\ 0, & \text{en el resto} \end{cases}$$

- (a) Calcula k para que $f_X(x)$ sea función de densidad.
 - (b) Calcula y representa la función de distribución, (suponga $k = 0,5$).
 - (c) Calcula la media.
 - (d) Calcula $p(X \geq 0,75)$.
7. La demanda semanal de cierta materia prima por parte de una empresa es de tipo aleatorio y tiene por función de densidad:

$$f_X(x) = \begin{cases} 1 - (x/2), & \text{si } 0 < x < 2 \\ 0, & \text{en el resto} \end{cases}$$

- (a) Comprueba que $f_X(x)$ sea función de densidad.
 - (b) Calcula y representa la función de distribución.
8. Dada la siguiente función de distribución de una variable aleatoria X

$$F_X(x) = \begin{cases} 0, & \text{si } x < 2 \\ \frac{(x-2)^3}{8}, & \text{si } 2 \leq x < 4 \\ 1, & \text{si } x \geq 4 \end{cases}$$

Se pide:

(a) Obtén la función de densidad.

(b) Obtén $p(1 < x \leq 3)$, $p(x > 3)$, $p(x \leq 2,5)$.

9. Dada la siguiente función de distribución de una variable aleatoria X

$$F_X(x) = \begin{cases} 0, & \text{si } x < 0 \\ x^3, & \text{si } 0 \leq x < 1 \\ 1, & \text{si } x \geq 1 \end{cases}$$

(a) Obtén la función de densidad.

(b) Obtén su media.

10. Sea una variable aleatoria continua con la siguiente función de densidad:

$$f_X(x) = \begin{cases} \frac{1}{6}, & \text{si } 1 \leq x < 2 \\ \frac{1}{3}, & \text{si } 3 \leq x < 4 \\ \frac{1}{2}, & \text{si } 5 \leq x < 6 \\ 0, & \text{en otro caso} \end{cases}$$

(a) Calcula la función de distribución.

(b) Calcula la esperanza y la varianza.

(c) Calcula $p(x \leq 1,75)$.

11. De una determinada variable aleatoria X se conoce que $p(X = 3) = 0,5$ y que su función de distribución viene dada por:

$$F_X(x) = \begin{cases} 0, & \text{si } x < 0 \\ \frac{1}{3}, & \text{si } 0 \leq x < 1 \\ \frac{5}{12}, & \text{si } 1 \leq x < 2 \\ \frac{5}{12} + k, & \text{si } 2 \leq x < 3 \\ 1, & \text{si } x \geq 3 \end{cases}$$

- (a) Determina el valor de k para que $F_x(x)$ sea función de distribución de la variable X .
 - (b) Determina la función de masa de probabilidad de dicha variable.
 - (c) Calcula la $p(E[X] < X \leq 3)$.
 - (d) Calcula la $Var[X]$.
12. El departamento de control de calidad de una empresa que fabrica pañuelos sabe que el 5 % de los mismos tienen algún tipo de defecto. Los pañuelos se empaquetan en cajas con 15 pañuelos. Calcula la probabilidad de que una caja contenga:
- (a) Dos pañuelos defectuosos.
 - (b) Menos de tres pañuelos defectuosos.
 - (c) El número esperado de pañuelos defectuosos de una caja.
 - (d) Si una empresa empaqueta esas cajas a su vez en lotes que contienen 150 pañuelos. Calcule la probabilidad de que el número de pañuelos del lote esté comprendido entre 6 y 9.
13. Se envían 15 invitaciones a los alumnos de primero de Gestión y Administración Pública para asistir a una conferencia. De experiencias anteriores se sabe que la probabilidad de que un alumno acepte la invitación es de 0.8. Calcula la probabilidad de que:
- (a) Como máximo 12 estudiantes acepten la invitación.
 - (b) Acepten la invitación más de 6 y menos de 10.
 - (c) Para otra conferencia, se decidió enviar la invitación a 90 alumnos de primer curso de la citada titulación. ¿Cuál es el número de alumnos que se espera acudan a la conferencia?
14. Una cadena de televisión privada concede un crédito anual a las personas que deciden abonarse a su programación. Por experiencias anteriores se sabe que el 10 % de las personas a las que se les concede el crédito no realizan finalmente el pago. En un período de promoción, y de manera independiente, han decidido abonarse a la programación de la cadena 13 clientes; por cada uno que finalmente realiza el pago, la compañía obtendrá un ingreso de 33 500 pesetas.
- (a) Calcula la probabilidad de que al menos 5 personas no realice el pago correspondiente.
 - (b) Calcula el ingreso esperado que tendrá esa compañía con esos 13 clientes.
 - (c) Si durante otra promoción han decidido solicitar el crédito 42 clientes, ¿cuál es la probabilidad de que entre ellos haya como mucho 10 que finalmente no paguen?

15. En una compañía que se dedica a la fabricación de impresoras se ha detectado un fallo en el proceso de fabricación que ha provocado que el 10 % de las unidades fabricadas sean defectuosas. La compañía ha distribuido las impresoras a los centros en lotes de 10 unidades:
- (a) Calcula el porcentaje de lotes que tienen alguna impresora defectuosa.
 - (b) Se han repartido 150 lotes. Cada lote será devuelto si se encuentran más de dos impresoras defectuosas. Calcula la probabilidad de que sean devueltas más de 15 lotes.
 - (c) ¿Cuál es el número de lotes que se esperan que sean devueltos?
16. El número de personas que llegan en una hora a una ventanilla de una oficina de la Agencia Tributaria, sigue una distribución de Poisson con media 7. Si acuden a la ventanilla más de 5 personas en una hora, se forma cola.
- (a) ¿Cuál es la probabilidad de que se forme cola en una ventanilla en una hora determinada?
 - (b) En cada oficina hay 10 ventanillas independientes, ¿cuál es la probabilidad de que en una hora se forme cola en 7 de esas 10 ventanillas?
 - (c) Si la Agencia Tributaria tiene 210 oficinas independientes en todo el país, ¿cuál es la probabilidad de que en una determinada hora haya al menos 1 450 ventanillas con cola?
 - (d) Calcula la probabilidad de que lleguen como máximo 75 personas a una determinada oficina en una hora.
17. El servicio de urgencias de la Clínica ASISA de Jerez recibe un promedio de 5 enfermos por hora.
- (a) Calcula la probabilidad de que tres enfermos acudan en una hora seleccionada al azar.
 - (b) Calcula la probabilidad de que menos de tres enfermos acudan en una hora seleccionada al azar.
 - (c) Si en el servicio de urgencias del Hospital del S.A.S. se reciben un promedio de 20 enfermos por hora. Calcula, aproximadamente, la probabilidad de que el número de enfermos que acudan en una hora esté comprendido entre 18 y 22, inclusive.
 - (d) Calcula la probabilidad de que se colapsen las urgencias del Hospital del S.A.S. en una hora determinada, si dicho hecho ocurre si se reciben más de 30 enfermos en una hora.
18. De un estudio sobre las sucursales de la entidad bancaria se sabe que el número diario de descubiertos en cada una de ellas sigue una distribución de Poisson de varianza 15.

- (a) Calcula la probabilidad de que en un día determinado se produzcan menos de 10 descubiertos.
 - (b) ¿Cuál es la probabilidad de que un día determinado, en 2 sucursales independientes, se produzcan en total más de 34 descubiertos?
 - (c) ¿Cuál es la probabilidad de que en un día determinado, en ninguna sucursal de las 250 que la entidad tiene en cierta región, se hayan producido más de 25 descubiertos?
19. El índice de cotización diaria de ciertas acciones en la Bolsa de Tokio se distribuye normalmente con media 500 puntos. Se sabe que un inversor decide invertir un día si el índice se sitúa entre 525 y 575 puntos y que el 28,1 % de los días el índice de cotización es inferior a 485,5 puntos. Calcula:
- (a) La desviación típica de la cotización diaria de las acciones.
 - (b) La probabilidad de que un día cualquiera un inversor invierta en dichas acciones.
 - (c) La probabilidad de que, de 200 días, invierta a lo sumo en 31 de ellos.
 - (d) El número medio de días al mes (20 días) en los que se espera que la cotización supere los 475 puntos.
20. En cierta ciudad la cuota por contribuyente del impuesto de vehículos de tracción mecánica sigue una distribución Normal con media 5 000 unidades monetarias. El porcentaje de contribuyentes que pagan menos de 3 000 u.m. es de 15,87 %.
- (a) Calcula el porcentaje de contribuyentes que paga una cuota comprendida entre 2 500 y 4 000 u.m.
 - (b) En la misma ciudad la cuota por contribuyente del impuesto de servicio de recogida de basuras se distribuye también normalmente con media 5 000 u.m. y desviación típica 1 500 u.m. Si las cuotas de cada impuesto son independientes:
 - (i) ¿Qué distribución sigue la cuota total por contribuyente procedente de la suma de las cuotas de estos dos impuestos?
 - (ii) ¿Cuál es el porcentaje de contribuyentes que paga una cuota total superior a las 8 000 u.m.?
 - (iii) Si se seleccionan al azar 120 contribuyentes de dicha ciudad, calcula la probabilidad de que al menos 90 de ellos paguen una cuota total superior a 8 000 u.m.
21. Los alumnos de Estadística Administrativa de la Facultad de Ciencias Sociales y de la Comunicación de Jerez han realizado un estudio sobre la estatura de los alumnos matriculados en el centro. Obtuvieron que el 8,38 % de los alumnos tienen una altura inferior a 160 cm y que el 10,38 % tienen

una altura superior a 185 cm. Suponiendo que las alturas siguen una distribución Normal:

- (a) Calcula la estatura media de los alumnos de la Facultad señalada y su desviación típica.

Para el apartado siguiente considérese $\mu = 173$ cm y $\sigma = 9,5$ cm.

- (b) Se estima que la citada Facultad tiene 1 500 alumnos. A los alumnos con una altura superior a 190 cm se les propondrá realizar una prueba para entrar en el equipo de baloncesto. Calcula la probabilidad de que fuesen seleccionados al menos 40 alumnos para realizar dicha prueba.

22. El peso de cierto artículo se distribuye normalmente. Se sabe que el 20,05 % de los artículos tienen un peso superior a 912 gramos y que el 69,85 % pesan menos de 800 gramos.

- (a) Calcula la media y la desviación típica de la distribución.
- (b) ¿Qué porcentaje de los artículos tienen peso inferior a 700 gramos?
¿Qué porcentaje de los artículos tienen peso comprendido entre 500 y 736 gramos?
- (c) Los citados artículos se empaquetan en grupos de diez. Determina la distribución del peso de un paquete, suponiendo independencia en los pesos de los artículos.
- (d) Calcula la probabilidad de que de los diez artículos que forman un paquete, exactamente tres pesen más de 700 gramos.

Recursos

Seguidamente exponemos una serie de páginas de la web donde podemos encontrar algunos recursos electrónicos relacionados con esta parte:

- (a) *[http : //psych.colorado.edu/ ~ mcclella/java/normal/handleNormal.html](http://psych.colorado.edu/~mcclella/java/normal/handleNormal.html)* muestra las probabilidades bajo la zona sombreada de la campana.
- (b) *[http : //www.ruf.rice.edu/ ~ lane/stat_sim/normal_approx/index.html](http://www.ruf.rice.edu/~lane/stat_sim/normal_approx/index.html)* de la Universidad de Rice (Tejas), muestra una distribución Binomial con parámetros que se pueden elegir y la compara con la distribución Normal. También calcula las probabilidades para un intervalo y las compara con la aproximación Normal.

Parte VI

**MÉTODOS DE
INFERENCIA
ESTADÍSTICA**

Capítulo 18

Introducción al Muestreo

18.1. Definiciones

Definición 18.1 *Llamamos Población o Universo a un conjunto de individuos o unidades del que se requiere información.*

Definición 18.2 *Al número de unidades que forman parte de la población lo llamamos tamaño poblacional.*

Definición 18.3 *Cuando el investigador requiere información de todos y cada uno de los elementos de la población estadística se dice que se está realizando un censo.*

Ventaja del censo

El censo posee una gran precisión.

Inconvenientes del censo

- (a) Tienen un costo elevado, en tiempo y dinero, si el número de elementos de la población es grande.
- (b) La obtención de la información es lenta. Puede ocurrir que la característica a estudiar haya variado en el transcurso de la realización del censo.
- (c) La exhaustividad del censo influye a veces en que la calidad de la información que se obtiene es baja.
- (d) Posible transformación o destrucción de los elementos de la población.

Esto hace necesario buscar alguna solución que resuelva en todo o en parte los problemas antes enunciados pero que siga permitiendo obtener conclusiones

sobre la población. Es en este camino donde interviene la muestra, el muestreo, la encuesta y más generalmente las técnicas de la Inferencia Estadística.

Definición 18.4 *Llamamos muestra a cualquier subconjunto representativo de la población.*

Definición 18.5 *Al número de unidades que compone la muestra lo llamamos tamaño muestral.*

Definición 18.6 *Cuando el investigador requiere información de sólo unos cuantos elementos de la población se dice que se está realizando una encuesta.*

Definición 18.7 *Muestreo es el proceso de selección de una muestra de la población.*

Definición 18.8 *A la metodología concerniente a hacer predicciones sobre la población estadística basándose en la información contenida en la muestra recibe el nombre de Inferencia Estadística.*

En primer lugar definiremos algunos conceptos básicos relacionados con la Inferencia Estadística.

Definición 18.9 *Como los valores de una muestra son aleatorios (a muestras distintas tendremos valores distintos), una muestra puede considerarse como n variables aleatorias $(X_1, X_2, \dots, X_n)$ independientes e idénticamente distribuidas.*

Definición 18.10 *Al conjunto de valores observados $(x_1, x_2, \dots, x_n)$ se llama realización de la muestra.*

Definición 18.11 *Una función $g(X_1, X_2, \dots, X_n)$ de la muestra se llama estadístico.*

Definición 18.12 *Sea $f(X; \vartheta)$ la función de densidad de X , y ϑ un parámetro desconocido, se llama estimador de ϑ a una variable aleatoria función de la muestra $\hat{\vartheta} = g(X_1, X_2, \dots, X_n)$ que en algún sentido que precisaremos, alcanza valores próximos al valor desconocido de ϑ .*

EJEMPLO 18.1

(a) *Media muestral:*

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

(b) Varianza muestral:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

(c) Cuasi-varianza muestral:

$$S_c^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Definición 18.13 Se llama estimación a uno de los posibles valores que puede tomar el estimador al obtener una realización muestral particular

$$\hat{\vartheta}_0 = g(x_1, x_2, \dots, x_n).$$

OBSERVACIÓN 18.1 La estimación es un número mientras el estimador es una variable aleatoria.

EJEMPLO 18.2

Población	Párametro	Estadístico	Estimación
Empleados de una fábrica	Rotación media del personal por año	Rotación media del personal en un período de 1 mes	Rotación del personal de un 9 % al año
Solicitantes del puesto de gerente	Media de la educación en años	Media de la educación de cada 5 solicitantes	18 años de educación
Chicos de una comunidad	Proporción que tiene antecedentes penales	Proporción en una muestra de 50 chicos con antecedentes	2 % tiene antecedentes

La Estadística Inferencial consiste fundamentalmente en la resolución de dos grandes categorías de problemas:

(a) La Estimación.

Consiste en determinar el valor del parámetro poblacional desconocido, de dos formas posibles:

- (i) Estimación puntual: se produce la determinación de un valor concreto para el parámetro poblacional.
- (ii) Estimación por intervalos: se produce la determinación de un intervalo en el que quede incluido el valor del parámetro con cierto grado de probabilidad.

(b) El Contraste de Hipótesis.

Consiste en determinar si es aceptable, a partir de los datos muestrales, que el parámetro poblacional tome un valor determinado o pertenezca a un intervalo determinado.

18.2. Introducción a la Teoría de Muestras

Definición 18.14 *Una vez visto el objetivo y la necesidad de la realización de una encuesta, hay que preparar un plan o método a seguir llamado plan de la encuesta. Alguno de sus pasos son:*

- (a) *Sopesar si los resultados compensan los gastos.*
- (b) *Averiguar cuál es la mejor fuente de información.*
- (c) *Identificar el grupo declarante.*
- (d) *Reunir listados actualizados de declarantes y de direcciones, lo que se entiende por **marco de la encuesta**.*
- (e) *Elegir el momento adecuado en que debe pasarse la encuesta.*
- (f) *Establecer el método adecuado de recogida de datos, que puede ser:*
 - (i) *La encuesta personal clásica sobre papel (PAPI). La entrevista personal es un método caro y puede introducir sesgos por parte del entrevistador, aunque se obtiene una alta tasa de respuesta.*
 - (ii) *La encuesta por correo. Tiene un escaso índice de respuestas aunque es un método barato.*
 - (iii) *La entrevista telefónica asistida por ordenador (CATI). Este tipo de encuesta tiene un bajo coste y una alta tasa de respuestas, pero la entrevista ha de tener una duración limitada.*
 - (iv) *Las entrevistas personales asistidas por ordenador (CAPI).*
 - (v) *Las auto-entrevistas asistidas por ordenador (CADI).*
- (g) *Asegurar el anonimato de los declarantes.*
- (h) *Pedir datos disponibles.*
- (i) *En el cuestionario utilizar terminología y definiciones uniformes y claras. Las preguntas deben ser lo más neutras posibles. Preocupada por la influencia del estilo de las preguntas, una empresa decidió preguntar la misma cuestión de dos formas distintas. Se preguntaba a los entrevistados qué programas deberían sacrificarse en caso de que se produjera un recorte presupuestario. En la lista de opciones se encontraba "Ayuda a los necesitados".*

Esta opción fue elegida por el 7% de los entrevistados. Dos meses después, se formuló la misma pregunta, se sustituyó la anterior opción por "Asistencia social". En esta ocasión el 39% de los entrevistados eligieron esa opción.

- (j) *Evitar instrucciones largas y excesos de detalles.*
- (k) *No influir en las respuestas.*
- (l) *Incluir preguntas para saber la veracidad de las respuestas anteriores. Por ejemplo, en las encuestas políticas, incluir la intención de voto y la valoración del líder.*
- (m) *Tener buenas relaciones con el público con objeto de incrementar el porcentaje de respuestas.*
- (n) *Efectuar un plan de ensayo para detectar errores.*

El riesgo principal de muestreo es obtener una muestra que no sea representativa de la población. Decimos entonces que la muestra está sesgada o tiene sesgos. Tipos de sesgos:

- (a) *Sesgo de selección.* Se produce cuando algunos miembros de la población tienen una probabilidad más alta que otros de estar representados en la muestra y ésta puede no ser representativa de la población.
En 1936, los encuestadores más importantes predijeron que Landon ganaría a Roosevelt mediante una amplia encuesta por teléfono y se equivocaron. El sesgo se originó porque los abonados al servicio telefónico pertenecían entonces a la clase media alta mayoritariamente.
- (b) *El sesgo de la no respuesta* puede hacer que la población observada no coincida con la población que se desea estudiar.

18.3. Muestreos no probabilísticos

Definición 18.15 *Muestreo no probabilístico o no aleatorio, cuando no puede calcularse de antemano cuál es la probabilidad de obtener cada una de las muestras que sea posible seleccionar. Para esto es necesario que la selección no pueda considerarse como un experimento aleatorio.*

Tipos de muestreos no probabilísticos:

- (a) *Muestreo opinático o intencional.* Es el investigador el que determina cuáles son las unidades que van a formar parte de la muestra. En este sentido, el *método Delphi* es una técnica de predicción grupal, que tiene como objetivo obtener una opinión consensuada de un grupo de expertos acerca de un

tema en concreto. El estudio se realiza en varias rondas, consistentes en repartir el cuestionario varias veces junto a los resultados del anterior. Con esto se busca que los expertos recapaciten sus opiniones, adopten otras si les parecen más acertadas y converjan lo máximo posible. Obtenida la estabilidad de las respuestas o fijado un número máximo de iteraciones, finaliza el proceso con un informe a cargo del coordinador del grupo.

- (b) *Muestreo errático o sin norma.* La muestra se selecciona por razones de comodidad o capricho.

Estos tipos de muestreo abaratan la recogida de información pero carecen de rigor científico.

A modo de ejemplo, en 1948, los encuestadores más importantes predijeron el resultado de la elección presidencial en Estados Unidos entre Harry Truman, entonces presidente, y Thomas Dewey, gobernador de Nueva York, como favorable a Dewey. Incluso hubo periódicos que sin esperar al recuento de las urnas dieron como vencedor a Dewey. El error estuvo en que el método de muestreo que se utilizó fue no probabilístico.

18.4. Muestreos probabilísticos

Definición 18.16 *Muestreo probabilístico o aleatorio cuando puede calcularse de antemano cuál es la probabilidad de obtener cada una de las muestras que sea posible seleccionar. Para esto es necesario que la selección pueda considerarse como un experimento aleatorio.*

Definición 18.17 *Si todas esas posibles muestras tienen la misma probabilidad de ser escogidas, una cualquiera de esas muestras recibe el nombre de muestra aleatoria simple (m.a.s.) o sin reemplazamiento.*

Ventajas de estos tipos de muestreo

- (a) Son precisos.
- (b) Las conclusiones son generalizables, tienen rigor científico.

Inconveniente

- (a) Son más caros que los no probabilísticos.

18.4.1. Muestreo aleatorio simple

Supongamos una población finita de N elementos. La forma de extraer una muestra aleatoria simple (m.a.s.) o sin reemplazamiento de tamaño n es:

- (a) Se numeran todos los elementos de la población.

- (b) A través de un sorteo se escogen n sin reemplazamiento. Algunas veces el sorteo es sustituido por la elección de unos números de una *tabla de números aleatorios* (*Técnicas de Montecarlo*). Una tabla de números aleatorios es un conjunto de números dispuestos de manera que cada dígito tiene la misma probabilidad de aparecer. Podemos construir una tabla de números aleatorios poniendo los números premiados en los sorteos de la lotería uno detrás de otro. Se agrupan los números aleatorios en bloques de tantos dígitos como tenga N y tomar n números aleatorios de la tabla.

18.4.2. Muestreo aleatorio con reemplazamiento

Supongamos que un elemento puede estar más de una vez en la muestra. Después de escoger un elemento, lo devolvemos a la población. En este tipo de muestreo todos los elementos tienen la misma probabilidad de ser escogidos, pero no todas las muestras tienen las mismas probabilidades de ser escogidas.

OBSERVACIÓN 18.2 *Se puede demostrar que el muestreo aleatorio con reposición es menos preciso que el muestreo sin reposición, es decir, en el muestreo sin reposición se necesita menos tamaño de la muestra para cometer el mismo error que el caso del muestreo con reposición.*

18.4.3. Muestreo estratificado

Este muestreo está indicado cuando:

- (a) Existan subpoblaciones y los elementos sean heterogéneos respecto a la variable que se desea estudiar.
- (b) En el caso de agencias u organismos, dispongamos de sucursales en distintos puntos, cada uno de ellos supervisa la encuesta en su correspondiente estrato poblacional, con el consiguiente ahorro de costes de organización, desplazamientos, etc.

Definición 18.18 *La estratificación es un procedimiento que consiste en dividir la población en un número de subpoblaciones o estratos, no solapados, y después tomar una muestra de cada estrato.*

Definición 18.19 *Si los elementos seleccionados de cada estrato constituyen una m.a.s., el procedimiento completo se llama muestreo aleatorio simple estratificado.*

Ventajas de este tipo de muestreo

- (a) En el muestreo estratificado todos los estratos tienen representación en la muestra. Ésta podría reducirse a un sólo elemento por estrato, cuando la población, heterogénea en su conjunto, dispone de porciones o partes que internamente tienen una gran homogeneidad.

- (b) Un número moderado de estratos aumentan la precisión del m.a.s. y podemos obtener una estimación de la variable dentro de cada estrato con una precisión suficiente. Sin embargo, cuando el número de estratos se multiplica la precisión disminuye.
- (c) En principio, el proceso de muestreo se efectúa de forma independiente en cada estrato, lo que permite la aplicación simultánea de diversos tipos de muestreo, de acuerdo a la información de la que se tenga, y según sea el coste y las razones que motivaron la estratificación.

Definición 18.20 *Se llama afijación a la distribución de las n unidades entre los diversos estratos.*

Tipos de afijación:

- (a) *Afijación uniforme*, consiste en asignar el mismo número de unidades muestrales a cada estrato. Con este tipo de afijación se da más importancia a los estratos pequeños. Sólo es conveniente cuando todos los estratos tengan tamaños parecidos.
- (b) *Afijación proporcional*. Las n unidades de la muestra se distribuyen proporcionalmente a los tamaños de los estratos. El muestreo de afijación proporcional también se denomina *autoponderado*, pues se demuestra que todas las unidades de la población tienen la misma probabilidad de pertenecer a la muestra y conduce a cálculos fáciles.
- (c) *Afijación de mínima varianza o de Neymann*, si las aportaciones a la muestra de cada estrato se calculan de acuerdo al criterio de minimizar la varianza de los estimadores, es decir, maximizar la precisión de los estimadores.
- (d) *Afijación óptima*, si las aportaciones de cada estrato a la muestra se han calculado bajo el criterio de minimizar costes.

18.4.4. Muestreo por conglomerados unietápico

Los tipos de muestreo anteriores son sólo aplicables cuando disponemos de la relación de los elementos de la población, o de los posibles estratos. Si esto no es así, está especialmente indicado el de conglomerados.

La población se divide en partes, grupos o subpoblaciones similares y cuyos elementos son de naturaleza heterogénea dentro de estas subpoblaciones o *conglomerados*.

Población	Variable	Elementos	Conglomerado
Ciudad A	Gastos en ropa	Personas	Hogares
Ciudad B	Características hogares	Hogares	Edificios
Instituto	Planes de estudio	Estudiantes	Cursos
Población rural	Actitudes sociales	Adultos	Pueblos

Se elige una m.a.s de conglomerados y dentro de los elegidos se estudian todos los elementos. La situación ideal sería cuando un sólo conglomerado representara a toda la población.

Ventajas de este tipo de muestreo

- (a) Se suele perseguir una reducción en el costo de la formación de listados o de las visitas a unidades muy esparcidas. No hace falta un marco muestral completo, basta con elaborar un marco parcial. Sólo se requiere la lista de las unidades primarias (conglomerados), y sólo censar las unidades de los conglomerados que han sido escogidos.

Inconvenientes

- (a) En el muestreo por conglomerados a diferencia del estratificado no todos los conglomerados tienen una representación en la muestra. Luego si los conglomerados son heterogéneos entre sí, como sólo se analizan algunos de ellos, la muestra final puede no ser representativa.
- (b) Desde el punto de vista de la precisión, interesa que los conglomerados sean homogéneos entre sí y heterogéneos dentro, al revés que con los estratos. En general, suele ser menos preciso que el m.a.s., ya que las unidades del interior de los conglomerados tienden a parecerse.

18.4.5. Muestreo de conglomerados con submuestreo

Este tipo de muestreo, también llamado bietápico, está especialmente indicado si en los n conglomerados existen una cierta homogeneidad interna, en vez de estudiar todos los elementos de los conglomerados seleccionados, consideramos los conglomerados como unidades primarias o de primera etapa, y obtener muestras dentro de cada conglomerado en una segunda etapa. Así por ejemplo:

Unidades primarias	Unidades secundarias
Hogares	Personas
Edificios	Hogares
Cursos	Estudiantes
Pueblos	Adultos

Ventajas e inconvenientes de este tipo de muestreo

Respecto al muestreo unietápico tiene la ventaja de extender mejor la muestra en la población, es decir, tomar un mayor número de unidades primarias a cambio de observar solamente una fracción de unidades secundarias. Por contra, al aumentar el número de conglomerados o unidades primarias se encarece el

muestreo. En la práctica hay que buscar una solución de compromiso entre los requerimientos de precisión y de coste.

OBSERVACIÓN 18.3 *Un tipo de muestreo interesante es el llamado **muestreo adaptativo**. En éste, el procedimiento para seleccionar las unidades que serán incluidas en la muestra puede depender de los valores de la variable de interés observada durante la encuesta.*

Si una unidad satisface la condición, el resto de los elementos del entorno también se añade a la muestra. Estos elementos constituyen un conglomerado.

18.4.6. Muestreo sistemático

Cuando los elementos de la población están ordenados en listas, una alternativa más fácil de ejecutar que el muestreo aleatorio simple es el muestreo sistemático.

Supongamos que el número de elementos de la población es divisible entre el número de unidades de la muestra $\frac{N}{n} = k$. Podemos considerar la población dividida en n zonas de k unidades. Selecciono al azar una unidad en la primera zona con probabilidad $\frac{1}{k}$. Las $n - 1$ unidades restantes de la muestra son las que ocupan el mismo lugar relativo en las restantes zonas.

Ventajas de este tipo de muestreo

- (a) Es de fácil aplicación y comprobación.
- (b) Las muestras se extienden de forma más regular, si los elementos más cercanos son más semejantes que los alejados.
- (c) Si los elementos en la lista están ordenados al azar, este procedimiento, es equivalente al muestreo aleatorio simple.
- (d) Puede ser considerado como un caso particular del muestreo por conglomerados dónde sólo es escogido uno.
- (e) Recoge el posible efecto de la estratificación.
- (f) En el muestreo sistemático, todos los grupos de unidades tienen una unidad perteneciente a cada zona o grupo en que hemos dividido la población.

Inconvenientes

- (a) La presencia de periodicidades ocultas.
- (b) En realidad sólo hay una elección al azar, la del primer elemento.

18.4.7. Muestreo bifásico

Se emplea en aquellas ocasiones en las que, antes de llevar a cabo la muestra, es necesario obtener una información adicional, mediante otra muestra. En general, la muestra de la primera fase o preliminar se elige de mayor tamaño y de menor coste que la de la segunda. Mientras que en muestreo bifásico se utilizan las mismas unidades de muestreo en todas las fases, en el muestreo polietápico existía una jerarquía de unidades de muestreo y éstas varían con las etapas.

EJEMPLO 18.3 *La siguiente tabla recoge los ingresos netos mensuales, en miles de pesetas, ordenados de izquierda a derecha, de 70 personas de Jerez.*

140	165	85	180	195	70	205	220	135	97	140	79	112	310
175	148	302	82	89	278	375	99	137	241	111	131	66	191
211	77	177	166	306	133	117	287	95	287	300	158	209	128
169	87	304	77	125	86	307	288	200	123	192	321	121	176
213	166	237	88	304	158	95	230	224	58	400	245	161	59

Comenta como estimarías la proporción de personas que tienen ingresos superiores a 182 000 pesetas al mes, si dicha estimación se basa en:

- En una muestra aleatoria de tamaño 10.*
- En una muestra estratificada de tamaño 5 de afijación proporcional, si la primera fila corresponde a mujeres y las otras cuatro a hombres.*
- En una muestra por conglomerados de tamaño 1, si las cuatro primeras filas pertenecen al barrio de S. Telmo y la última al de Santiago.*
- En una muestra por conglomerados bietápico de tamaño 1 en primera etapa y 5 en segunda.*
- En una muestra sistemática de tamaño 5.*

Solución:

- Utilizando las dos primeras columnas de la tabla de números aleatorios, selecciono los ingresos de los 10 primeros números menores que 70, la estimación de la proporción personas que cumplen la condición pedida en la población se consigue a través de la proporción de personas que en la muestra cumplen la citada condición.*
- La primera fila es el estrato de las mujeres y las otras cuatro es el estrato de hombres. Al ser la afijación proporcional debo seleccionar al azar una mujer del estrato de las mujeres y cuatro hombres dentro de su estrato.*
- Tendríamos dos conglomerados, S. Telmo y Santiago. Hago un sorteo entre los dos para ver cuál es el elegido y ahí encuesto a todos sus elementos.*

- (d) A diferencia del apartado anterior, una vez seleccionado el conglomerado, sólo encuesto a 5 elementos del conglomerado que he escogido.
- (e) Cada fila sería una subpoblación. Supongamos que el elemento elegido al azar de la primera subpoblación es el 11, luego los que entran a formar parte de la muestra será el 25, 39, 53 y 67. Posteriormente se actúa igual que en los anteriores apartados.

18.5. Otros tipos de muestreo

- (a) A veces la estratificación no es posible, o es muy cara, y se recurre en su lugar al *muestreo por cuotas*, donde se eligen las unidades de muestreo hasta completar una cuota (determinada previamente), proporcional al tamaño de las subpoblaciones del colectivo. El inconveniente de este muestreo es que, como los elementos de la muestra no están seleccionados aleatoriamente dentro de cada estrato, es fácil incurrir en sesgos de selección. Sin embargo, a pesar de estas limitaciones se utiliza mucho por su bajo coste. Combinado con otros métodos aleatorios no suele dar malos resultados.
- (b) Otro tipo de muestreo que se utiliza mucho cuando disponemos de una lista de los miembros de la población es el de las *rutas aleatorias*. En éste método se fija al entrevistador un recorrido al azar a partir de un punto para elegir la vivienda de la persona a entrevistar. La clave de este método está en asegurarse que los puntos iniciales de búsqueda son aleatorios, ya que si no, la muestra puede no serlo.

18.6. Métodos muestrales en el tiempo

La recogida de datos en varios períodos de tiempo, puede realizarse siguiendo distintos diseños muestrales:

- (a) Puedo utilizar la misma muestra, denominada *panel*, en ambas ocasiones.
- (b) Mantener en la segunda ocasión b unidades de la primera muestra, eliminar $n - b$ y añadir $n - b$ nuevas, *muestras de panel rotante*.
- (c) Utilizar una encuesta periódica, *muestreo repetido*, es decir, en la segunda ocasión obtener una muestra independiente de la primera.

Cada uno de los tres diseños tiene ventajas e inconvenientes.

La principal ventaja de las encuestas periódicas es la posibilidad de acumular información para estimar con mayor fiabilidad características. Esta cualidad también la presentan los paneles rotantes en mayor o menor medida.

Los trabajos de campo son más fáciles en un panel rotante, el panel puro necesita una gran dedicación para conseguir la colaboración continuada, así como conseguir la colaboración por primera vez en las encuestas periódicas.

Si queremos conocer los cambios individuales entre dos ocasiones temporales, las muestras independientes no son adecuadas, las mejores para este cometido son las muestras panel.

Capítulo 19

La Administración y las Estadísticas

19.1. El sistema estadístico en las Administraciones

El artículo 149 de la Constitución Española fija al Estado la competencia exclusiva en Estadística para fines estatales, teniendo las comunidades autónomas la competencia sobre las estadísticas para fines no estatales. La Ley 12/1989, de 9 de Mayo, de la Función Estadística Pública, es la norma legal básica para el ejercicio de la Estadística en la Administración General del Estado. La labor estadística corre a cargo del *Instituto Nacional de Estadística (INE)*, de los ministerios y del Banco de España.

De todas las operaciones estadísticas, aproximadamente el 25 % las realiza el INE y el 75 % restante corre a cargo de los ministerios. No todas esas operaciones se realizan mediante extracción de datos directos, sino que la gran mayoría utilizan datos de actos administrativos.

La Ley de 31 de Diciembre de 1945 crea el Instituto Nacional de Estadística, centro dedicado a la observación y estudios de los fenómenos colectivos de la vida española. También se crea el Consejo Superior de Estadística, órgano consultivo de los servicios estadístico estatales.

Actualmente, existen también una serie de comités, como el Interministerial e Interterritorial de Estadística, para coordinar con los ministerios y con las comunidades autónomas todas las operaciones estadísticas.

En la Ley 12/1989 se fija la obligatoriedad de facilitar los datos al Instituto y la obligación que tiene el INE de facilitarlos sólo en forma numérica, sin referencia alguna de carácter individual (*secreto estadístico*).

El *Instituto de Estadística de Andalucía (IEA)*, es el órgano director y coor-

dinador de la actividad estadística de la Junta de Andalucía. Fue creado por Ley 4/1989, de 12 de Diciembre, como organismo autónomo de carácter administrativo, con el fin de dar cumplimiento a las competencias reconocidas en el artículo 12.34 del Estatuto de Autonomía, el cual atribuye a la Junta de Andalucía la competencia plena en la realización de estadísticas para fines de la Comunidad Autónoma. Entre las actividades que realiza el IEA destacan el Índice de Producción Industrial de Andalucía (IPIAN), la Encuesta Industrial, las Tablas Input-Output de Andalucía y la Encuesta Coyuntural del Sector Turístico.

En el extranjero destacan la Oficina de Estadística de la Unión Europea (EUROSTAT), creada en 1953 para responder a las necesidades de la Comunidad del Acero y del Carbón. Es el principal órgano estadístico de la Comisión, centralizando la mayor parte de la actividad estadística y ejerciendo una función coordinadora. Lo mismo podemos decir de la Oficina Estadística de las Naciones Unidas (UNSTAT), que propone marcos conceptuales para las estadísticas en el ámbito mundial.

19.2. Las estadísticas demográficas

19.2.1. Censos y Padrón

El objetivo general de los *censos* de población es el recuento de la población y el conocimiento de su estructura. Además los censos sirven de base para la planificación económica y social.

Naciones Unidas considera que las características esenciales de un censo de población son:

- (a) La enumeración individual de la población.
- (b) La universalidad dentro de un territorio definido.
- (c) La simultaneidad refiriéndose a un instante determinado.
- (d) La periodicidad que permita homogeneidad de series históricas.

El Censo que se realizó en el 2001 en España se apoyó fuertemente en el *Padrón Continuo*. Este sistema es más preciso, ya que las cifras del Padrón son de mayor precisión que las censales, y mediante este sistema se abarataron costes, al aprovechar la infraestructura de los ayuntamientos.

El Censo de Población tiene carácter estadístico y su objetivo es informar sobre el número y la distribución de las principales características sociales, respetando el *secreto estadístico*. La finalidad del Padrón es meramente administrativa, acreditar la residencia en un domicilio, y sus datos son de carácter nominal.

El Censo de Población es una foto fija que se realiza cada 10 años, mientras que el Padrón cotidianamente se realizan altas y bajas.

Los censos de población decenal se han venido realizando desde 1900. A partir de 1950 se realiza conjuntamente con el Censo de Población el Censo de

Viviendas y Edificios. Con éste censo se quiere conocer el número, distribución y características de viviendas y edificios en España. Se realizan los años acabados en uno y la fecha de referencia es el 1 de Noviembre para el que se celebró en el año 2001. Los censos de Población y Vivienda han suministrado tradicionalmente el marco para las encuestas por muestreo.

19.2.2. Otras encuestas demográficas

La Estadística de Variaciones Residenciales (EVR) se elabora con información procedente de las altas y bajas de los Padrones Municipales y sirve para cuantificar los movimientos migratorios que tienen lugar entre distintos puntos de origen y destino.

Las estadísticas de movilidad de la población quieren conocer las características de los viajes realizados, el motivo de los desplazamientos, la duración, el número de participantes, etc.

Con las estadísticas vitales o también llamadas Estadísticas del Movimiento Natural de la Población (matrimonios, defunciones y nacimientos), se quieren establecer estimaciones de la población. Tales datos se han venido recogiendo en los Registros Civiles.

19.3. Las estadísticas económicas

19.3.1. La Encuesta de Población Activa

Quizás la principal encuesta es la Encuesta de Población Activa (EPA). Con ella se quiere conocer resultados de la fuerza de trabajo, número de parados, ocupados, etc. Se realiza en España desde 1964.

La encuesta va dirigida al conjunto de la población que reside en viviendas familiares en España. Se excluyen por tanto personas que permanecen en hospitales, hoteles, cuarteles o conventos. En conjunto la población excluida en la encuesta no llega al uno por ciento.

Se muestrean más de 60 000 viviendas al trimestre, utilizando un muestreo bietápico de conglomerados con submuestreo y estratificación en las unidades de primera etapa, que son las secciones censales. Las de segunda etapa son las viviendas, las cuales se renuevan en una sexta parte para evitar el cansancio de las familias.

Definición 19.1 *Para la EPA una persona está parada cuando declara:*

- (a) *No haber trabajado en la semana de referencia ni siquiera una hora ni por cuenta propia ni ajena.*
- (b) *Buscar trabajo de una manera activa durante el mes precedente.*
- (c) *Estar disponible para trabajar en como mucho quince días.*

Definición 19.2 *Y una persona ocupada será aquella que:*

- (a) *Ha trabajado en la semana de referencia al menos una hora por cuenta propia o ajena.*
- (b) *O quienes teniendo trabajo han estado temporalmente ausentes del trabajo por vacaciones o enfermedad.*

La unión de los parados y ocupados conforman la población económicamente activa.

Recientemente, en la EPA se han introducido una serie de cambios metodológicos:

- (a) Se ha modificado la definición de parado, restringiéndose los métodos de búsqueda activa de trabajo. Esto ha conllevado una disminución de la tasa de paro.
- (b) Se han revisado las proyecciones de población que se usaba hasta la fecha, basadas en el censo de 1991. Esto es debido a que la inmigración real ha sido superior a la considerada en las hipótesis iniciales. Esto ha provocado un aumento en valor absoluto en todas las categorías del mercado laboral (parados, ocupados, inactivos).
- (c) Se ha actualizado la muestra de secciones con la información del Censo y del Padrón. Debido al desarrollo de zonas periféricas la muestra de secciones no reflejaba el conjunto de la población. Con esta reforma se aumenta el número de ocupados.
- (d) Se han reponderado los factores de elevación para evitar la falta de respuesta por ausencia del domicilio entre el grupo de personas entre 25 y 55 años, formado por hogares monoparentales o en los que los dos cónyuges trabajan. Gracias a ello, la pirámide de la EPA coincide con la del Censo y esto ha hecho aumentar el número de ocupados.

La subsecretaría del Ministerio de Trabajo y Asuntos Sociales realiza la Encuesta de Coyuntura Laboral (ECL), para estimar los efectivos laborales desde la óptica de la empresa: estimaciones de altas y bajas, duración del tiempo de trabajo, causas de absentismo laboral y marco laboral. Es una encuesta de periodicidad trimestral dirigida a 12 000 empresas. Estima las previsiones de empleo a corto y medio plazo.

19.3.2. La Encuesta de Presupuestos Familiares

Con la Encuesta de Presupuestos Familiares (EPF), son varios los objetivos que se quieren cubrir:

- (a) Conocer la estructura de gasto e ingresos de los hogares españoles.

- (b) Obtener las ponderaciones para el cálculo del Índice de Precios de Consumo (IPC).
- (c) Estimar el consumo privado para la Contabilidad Nacional.
- (d) Posibilitar análisis en distintos campos de preocupación social: pobreza, nutrición, etc.

La nueva Encuesta Continua de Presupuestos Familiares (ECPF), realizada desde 1997, viene a sustituir a las encuestas coyunturales y estructurales efectuadas hasta la fecha.

Con esta nueva encuesta de panel rotante, con velocidad de rotación de 1/8, se seleccionan mediante un muestreo bietápico estratificado en la primera etapa, 8 064 viviendas al trimestre. Se obtienen así, datos trimestrales y datos anuales como suma de estimaciones trimestrales.

19.3.3. El nuevo Índice de Precios de Consumo

Seguidamente estudiaremos las características del nuevo Índice de Precios de Consumo (IPC), con base el año 2001. El Índice de Precios de Consumo es un instrumento estadístico cuyo objetivo es medir la evolución y cambios del nivel de precios del conjunto los bienes y servicios consumidos por los hogares residentes en España.

Entre sus características principales tenemos:

- (a) *Período base*: Es aquel cuyos precios sirven de referencia para medir la evolución de los mismos durante el periodo de vigencia del sistema. Actualmente es el año 2001. Otros que fueron considerados períodos base fueron los años 1936, 1958, 1968, 1976, 1983 y 1992.
- (b) *Período de referencia de las ponderaciones*: Es el período durante el cual se desarrolla la ECPF (Encuesta Continua de Presupuestos Familiares), que proporciona la información básica sobre los gastos de las familias en bienes y servicios de consumo para la obtención de las ponderaciones. La ECPF actual se realizó entre 1-4-1999 y 31-3-2001.
- (c) *Período de referencia de los precios*: Es el período con cuyos precios se comparan los precios corrientes, es decir, el período elegido para el cálculo de los índices simples. Durante el año 2002 se hizo coincidir con el año 2001 y para años posteriores será el mes de diciembre del año inmediatamente anterior al considerado.
- (d) *Estrato de referencia*: Es el conjunto de la población residente en viviendas familiares en España. Su estructura de gasto sirve para la determinación cualitativa y cuantitativa de la llamada cesta de la compra.

- (e) *Cesta de la compra*: Es el conjunto de bienes y servicios para los que se recogen los precios mensualmente, y cuya evolución representa la de todos los precios de consumo de la economía. La selección se realiza según el gasto realizado por el estrato de referencia y a través de la ECPF. Se consideran 484 artículos que se agrupan en 12 grandes grupos.
- (f) *Fórmula utilizada*: Hasta ahora se utilizaba un índice tipo Laspeyres con base fija. Este sistema tiene la ventaja de la comparabilidad de una misma estructura durante el período de vigencia del sistema, sin embargo, tiene el inconveniente de la pérdida de la vigencia, conforme pasa el tiempo y evolucionan las pautas de consumo. El nuevo sistema utiliza la fórmula del *índice de Laspeyres encadenado*, que consiste en referir los precios de los artículos del período corriente a los precios del año inmediatamente anterior. La fórmula general es la siguiente:

$$I_{L/0,t}^c = \prod_{k=1}^t I_{L/k-1,k} = I_{L/0,1} \cdot I_{L/1,2} \cdots I_{L/t-1,t}$$

donde,

$I_{L/0,t}^c$ es el índice de Laspeyres encadenado del período t referido al período 0.

$I_{L/k-1,k}$ es el índice de Laspeyres del período k referido al período $k - 1$.

que se calcularían según la expresión:

$$I_{L/k-1,k} = \sum_i \frac{p_{i(k-1)} q_{i(k-1)}}{\sum_i p_{i(k-1)} q_{i(k-1)}} \cdot \frac{p_{ik}}{p_{i(k-1)}} = \sum_i W_i^{k-1} I_{k-1,k}^i$$

Los períodos intermedios considerados en el IPC-Base 2001 corresponden a los meses de diciembre de todos los años desde el 2002 hasta el año inmediatamente anterior al que hace referencia el índice. Por ejemplo,

$$I_{L/01,ENE04}^c = I_{L/01,DIC02} \cdot I_{L/DIC02,DIC03} \cdot I_{L/DIC03,ENE04}$$

Durante el año 2002, que se consideró de transición, los índices se calcularon siguiendo un procedimiento diferente.

- (g) *Ponderaciones*: Representan la importancia relativa que cada artículo de la cesta de la compra tiene frente a los demás. Los gastos de consumo se han clasificado según la COICOP (Classification Of Individual Consumption by

Porpuse). De ocho grupos anteriores el IPC pasa a 12. Las ponderaciones vienen dadas por la expresión:

$$W_i = \frac{\text{Gasto realizado en las parcelas representadas por el artículo } i}{\text{Gasto total}}$$

Las ponderaciones IPC-Base 2001 para los doce grandes grupos aparecen en la tabla siguiente:

Grupos	Ponderación
1. Alimentos y bebidas no alcohólicas	218.630
2. Bebidas alcohólicas y tabaco	32.170
3. Vestido y calzado	99.280
4. Vivienda	110.260
5. Menaje	63.571
6. Medicina	28.062
7. Transporte	155.760
8. Comunicaciones	25.729
9. Ocio y cultura	67.263
10. Enseñanza	17.444
11. Hoteles, cafés y restaurantes	112.708
12. Otros	69.124
	1 000

OBSERVACIÓN 19.1

- (a) *El nuevo IPC es más dinámico y flexible ya que se revisarán los datos anualmente incluyendo cualquier bien cuyo consumo sea significativo. El IPC se adapta a los cambios de mercado y de los hábitos de consumo en un plazo muy breve de tiempo. Se puede detectar la aparición de nuevos bienes y servicios en el mercado para su inclusión en el IPC, así como la desaparición de los que se consideren poco significativos, como ejemplo tenemos que ha sido eliminada la máquina de escribir y se ha incorporado la comida preparada.*
- (b) *Las ponderaciones se actualizan periódicamente. Se ponderará por gasto, en lugar de por el número de unidades, como se venía haciendo hasta la fecha.*
- (c) *Debido a los cambios introducidos, los índices calculados a partir de Enero de 2001 deberán ser enlazados con los anteriores base 1992. Cada agregación tendrá su propio coeficiente de enlace, por lo que los índices enlazados pierden la propiedad de la aditividad, es decir, el índice general enlazado no resulta de la suma ponderada de los índices enlazados de los grupos.*

- (d) Además los precios medios provinciales de todos los artículos incluidos en la cesta de la compra que sirven para el cálculo de los índices elementales se calcularán mediante medias geométricas, para eliminar en lo posible la influencia de los precios extremos, en lugar de la media aritmética.*
- (e) Se incorporan los artículos rebajados y los precios en oferta durante todos los meses del año. No se recogerán, en cambio, las ofertas destinadas solamente a una parte de la población.*
- (f) Si a un artículo le falta el precio, no se aplicará el criterio del último precio que se tenga (carry-forward), ya que pueden haber pasado meses, sino que se estimará a partir del resto de las tomas.*

Para la correcta medición de la evolución de los precios es preciso estimar en qué medida la variación observada del precio es debida al cambio en la calidad del producto y qué parte de ella es achacable al precio, independientemente de la calidad.

Para solucionar este problema de la medición de un cambio de precio, debido a un cambio en la calidad del producto, se utilizan varios métodos. Los más utilizados en el IPC son:

- (a) El método DELPHI, que consiste en buscar la información de expertos para estimar el cambio.
- (b) Los precios de las opciones, que analiza los elementos de los componentes del antiguo y del nuevo modelo, para establecer el coste de la diferencia entre ambos.
- (c) El precio de solapamiento, basado en el valor de la diferencia de precio en el período que están vigentes los precios de ambos artículos, el nuevo y el viejo.
- (d) La regresión hedónica, para ver cuál es el efecto en el día de hoy y en años anteriores de los cambios de calidad. Consiste, básicamente, en aplicar modelos de regresión que permitan obtener un ajuste del precio de referencia sobre la base de una serie de variables que determinan el precio del artículo.

Entre otras aplicaciones, el IPC se emplea como medida de la inflación; para revisar contratos de arrendamiento, pensiones, salarios; como deflactor y para la actualización de primas de seguros, etc.

19.3.4. El Índice de Consumo Armonizado

El Índice de Precios de Consumo Armonizado (IPCA), es un indicador estadístico cuyo objetivo es proporcionar una medida común de la inflación que permita comparaciones internacionales y examinar, así, el cumplimiento que de

esta materia exige el Tratado de Maastricht para la entrada en la Unión Monetaria.

El INE lo ha empezado a publicar a partir del 7 de Marzo de 1997. En un afán de armonizar todos los IPC nacionales, algunas partidas quedaron excluidas en primera instancia hasta alcanzar unos criterios comunes para su inclusión, como la Sanidad o la Educación.

La única diferencia entre el nuevo IPC y el IPCA, a partir de Enero de 2002, es la distinta cobertura de la población: las ponderaciones del IPCA se obtienen del gasto de residentes y no residentes realizado dentro del territorio español, mientras que las del IPC se calculan a partir del gasto realizado por los residentes independientemente de dónde se realice el gasto.

19.3.5. El Índice de Producción Industrial

El Índice de Producción Industrial (IPI), mide la evolución del valor añadido generado por las ramas industriales (extractivas, manufacturas y producción y distribución de gas, agua y electricidad), salvo a la construcción.

El valor añadido al coste de los factores puede calcularse, fundamentalmente, a partir del volumen de negocios (excluido el IVA), más la producción inmovilizada, menos las compras de bienes y servicios, menos otros impuestos. Este índice tiene por objeto:

- (a) Como indicador dinámico, debe medir en cada momento la evolución de la producción y no su magnitud.
- (b) Como indicador coyuntural (junto con otros indicadores relativos al comercio exterior, paro, etc.), debe medir los movimientos a corto plazo. Su periodicidad es mensual.
- (c) Como indicador de agregación, debe resumir en pocas cifras una considerable cantidad de información.
- (d) Como indicador del volumen de producción, debe reflejar las variaciones en la cantidad y calidad, no en los precios.

El IPI utiliza la metodología de Laspeyres con ponderaciones fijas para el año base, el último el del año 2000.

Se define el índice I_t correspondiente al período t como:

$$I_t = \sum_i I_{it} \omega_{i0} = \sum_i \frac{q_{it}}{q_{i0}} \frac{q_{i0} p_{i0}}{\sum_i q_{i0} p_{i0}} = \frac{\sum_i q_{it} p_{i0}}{\sum_i q_{i0} p_{i0}}$$

donde q_{it} y q_{i0} son las producciones en la rama de la actividad i en el período corriente y en el período base respectivamente, y p_{i0} los precios en el año base.

Las ponderaciones necesarias para construir el IPI se obtienen a partir de la Encuesta Industrial de Empresas y la Encuesta Anual Industrial de Productos. La diferencia fundamental entre una y otra radica en que en la primera la unidad de encuesta es la empresa, y en la segunda es el establecimiento industrial perteneciente a empresas con 20 o más personas ocupadas. Se vienen realizando desde 1993.

Dichas encuestas aportan anualmente la información necesaria para seleccionar los productos y establecimientos a los que se mandará el cuestionario. Se envían cada mes 9000 cuestionarios que proporcionan información de 990 productos, lo que supone el 90 % del total.

El *Índice de Producción Industrial de Andalucía (IPIAN)*, es un indicador de coyuntura cuyo objetivo es medir la evolución a corto plazo, concretamente de forma mensual, del Valor Añadido Bruto (VAB) del sector industrial andaluz. Ante la imposibilidad de medir mensualmente el VAB generado en la industria, se toma una variable intermedia como indicador: la producción física. Por lo tanto, el IPIAN no pretende medir exactamente la evolución del VAB industrial, sino únicamente mostrar la dirección e intensidad con que dicha variable se mueve. El principal factor de identificación de una unidad de producción como perteneciente al sector industrial de la Comunidad Autónoma de Andalucía es el mantenimiento de un establecimiento o centro de trabajo permanente en el que se desarrollen actividades integradas en el sector industrial. En el IPIAN-94 el número de productos seleccionados es de 203, que ocupan el 84 % del empleo industrial de Andalucía y representan el 91 % del valor de la producción industrial regional.

El INE también calcula el Índice de Cifras de Negocios (ICN) y el Índice de Entrada de Productos (IEP). El IEP es un indicador adelantado del IPI y éste a su vez del ICN.

19.3.6. El Índice de Precios Industriales

Con el Índice de Precios Industriales (IPRI), se quiere medir la evolución de los precios de los productos industriales mes a mes, salvo la construcción y los servicios, en la primera etapa de comercialización, es decir, el precio de venta a la salida de la fábrica, sin IVA ni gastos de transporte. A diferencia del precio de adquisición con IVA que es la base del IPC. Se excluyen también las ventas en el mercado exterior y las ventas de productos importados.

Este índice tiene diversos usos:

- (a) Como indicador para el análisis de los precios, detectando en sus inicios las presiones inflacionistas. Su estudio, junto al IPC, nos sirve para conocer el mecanismo de formación de los precios.
- (b) Como deflactor, al objeto de convertir datos en valor a precios corrientes a precios constantes, o para deflactar las series en valor del IPI.

La metodología utilizada es la de Laspeyres, considerando constante la estructura relativa de las cantidades vendidas.

Se define el índice I_t correspondiente al período t se define como:

$$I_t = \sum_i I_{it} \omega_{i0} = \sum_i \frac{p_{it}}{p_{i0}} \frac{p_{i0} q_{i0}}{\sum_i p_{i0} q_{i0}} = \frac{\sum_i p_{it} q_{i0}}{\sum_i p_{i0} q_{i0}}$$

donde p_{it} y p_{i0} son los precios en el período corriente y en el período base respectivamente, y q_{i0} las cantidades vendidas en el año base.

Las ponderaciones se han calculado de acuerdo a la importancia relativa de las ramas de actividad y de los productos en 1990, según la información que suministra la Encuesta Industrial. El actual índice involucra 228 ramas de actividad, 6 000 empresas informantes, 950 productos incluidos en la cesta y 20 000 datos primarios.

19.3.7. Otras encuestas económicas

Dentro de las encuestas económicas a las que dedicamos esta sección, ocupa un lugar relevante el *Censo Agrario*. El último fue el de 1999. Sirve para conocer la estructura agraria y a partir de él se puede confeccionar un directorio de explotaciones con sus características, como la superficie de la explotación, la existencia de maquinaria, irrigación, ganadería, etc. Se puede decir que la práctica de censos agrarios, entendida como investigación estadística exhaustiva, comienza en España en el año 1962. Se realizan cada 10 años por recomendación de la FAO.

El seguimiento de los salarios agrarios se realiza mensualmente desde el Ministerio de Agricultura y Pesca con la confección de unos índices. Los datos se consiguen a través de muestreo.

En 1989 se implanta la Encuesta de Salarios en la Industria y los Servicios para proporcionar información sobre los niveles y evolución de la ganancia media por hora, la ganancia media por trabajador y mes, y el número medio de horas trabajadas.

A partir del 2001, el INE sustituye esta encuesta por otra mejorada que amplía el ámbito poblacional, al incluir asalariados que ejercen su actividad en centros de 5 ó menos empleados, y el ámbito sectorial, al incluir secciones de Sanidad y Educación. El número de unidades muestrales es de 18 600. Por otra parte, se procederá a contabilizar de manera real el número de días de vacaciones.

El coste laboral se define como el coste que incurre el empleador por la utilización del factor trabajo. El *Índice de Costes Laborales (ICL)* es una operación estadística continua, de carácter coyuntural y periodicidad trimestral, que informa sobre el coste laboral medio por trabajador y mes, coste laboral medio por hora efectiva trabajada y el tiempo trabajado y no trabajado.

La finalidad del ICL es hacer un seguimiento trimestral de este coste y de sus principales componentes. El ICL es un índice simple de variación del coste medio laboral. La base es la del año 2000.

La encuesta se extiende al conjunto de la construcción, industria y servicios. El marco poblacional es el Directorio de Cuentas de Cotización a la Seguridad Social. El ámbito poblacional comprende a todos los trabajadores por cuenta ajena asociados a cuentas de cotización.

El tipo de muestreo es el aleatorio estratificado con afijación óptima, en que las unidades muestrales son las cuentas de cotización. El criterio de estratificación se realiza atendiendo a tres variables: la comunidad autónoma, la actividad económica y el tamaño de las unidades.

La muestra se divide en cinco grupos de rotación de manera que en el primer trimestre de cada año se reemplaza el grupo con mayor antigüedad.

Debido a su importancia, la construcción recibe un tratamiento especial. El Ministerio de Fomento realiza diversas encuestas para conocer los grandes agregados que intervienen en la industria de la construcción. Mediante una serie de indicadores se quiere conocer la evolución del coste de los materiales, de la mano de obra, del coste de los pisos, etc.

Acabamos la sección de las estadísticas económicas comentando que el INE elabora un Sistema de Cuentas Regional, que resume en un conjunto de tablas y cuadros, con el fin de dar una visión comparable de la actividad económica de las regiones de España.

El Instituto de Estadística de Andalucía (IEA) elabora el marco input-output. Este es un instrumento estadístico-contable en el que se representan la totalidad de las operaciones de producción y de distribución que tuvieron lugar en Andalucía. Este trabajo adopta la metodología del nuevo Sistema Europeo de Cuentas (SEC-95).

19.4. Las estadísticas sociales

19.4.1. El Panel de Hogares

Una de las estadísticas sociales más importantes es el Panel de Hogares de la Unión Europea (PHOGUE). Su realización corre a cargo del INE, aunque se realiza a petición de EUROSTAT.

Su objetivo es conocer la evolución en el tiempo del nivel de vida, ingresos, sistema de protección, empleo, pensiones, nivel de formación, movilidad y condiciones de los hogares de la Unión Europea. El número de hogares entrevistados en la U.E. es de 76 500 de los 8 000 corresponden a España, siendo el número de personas entrevistadas en la Comunidad Económica Europea de 155 000. La recogida de la información se efectúa anualmente desde 1994 (primer ciclo), habiéndose completado 7 ciclos más.

Al ser una encuesta de tipo panel, las personas elegidas se mantuvieron en la muestra durante los siete años de duración de la investigación.

En el 2003 el PHOGUE ha sido sustituido por una nueva encuesta de rentas y condiciones de vida (EU-SILC).

19.4.2. Encuestas Turísticas

Los principales estudios de demanda turística son realizados por el Instituto de Estudios Turísticos (IET), adscrito a la Secretaría General de Turismo del Ministerio de Economía, mediante dos encuestas: La *Encuesta de Movimiento Turístico de los Españoles (FAMILITUR)*, que trata de conocer la cuantía y las características de los viajes realizados por la población residente en España y la *Encuesta de Movimiento Turístico en Fronteras (FRONTUR)*, dando una estimación del número de visitantes extranjeros que entran en España por distintas vías, la duración de la visita y el destino turístico.

Recientemente, en Enero de 2002, se realizó la *Encuesta de Gasto Turístico (EGATUR)*, realizada por el INE, el IET y el Banco de España, para estimar el gasto de los visitantes y el gasto de los españoles en sus salidas al extranjero. A través de 86 000 encuestas anuales se conoce el impacto del turismo en nuestra economía.

El estudio de la oferta turística (empresas ligadas a esta actividad), la realiza el INE a través de la Encuesta Estructural de las Empresas Turísticas. Con ellas obtenemos el número de empresas, el personal ocupado, gastos, ingresos, etc.

De la Encuesta de Ocupación Hotelera que realiza el INE surge el *Índice de Precios Hotelero (IPH)*. Es una medida estadística de la evolución del conjunto de precios aplicados por los empresarios a sus clientes: familias, empresas y tour-operadores. Mide, por tanto, la evolución de los precios del sector desde la óptica de la oferta.

El precio pagado por los hogares o familias que aparece recogido como un componente en el IPC, desde el lado de la demanda, no coincide con el que recibe el hotelero.

El IPH es un índice de tipo Laspeyres. Se puede definir como la relación entre los ingresos obtenidos en el período objeto y el período base, considerando el mismo número de pernотaciones en los dos periodos:

$$I_A^{mt} = \frac{\sum_{i=1}^6 P_i^{mt} N_i^c}{\sum_{i=1}^6 P_i^0 N_i^c}$$

donde I_A^{mt} es el índice agregado para el mes m del año t , P_i^{mt} es el precio conseguido con la tarifa i -ésima en el mes m y año t , P_i^0 es el precio conseguido con la tarifa i -ésima en el período base y N_i^c el número de plazas ocupadas en el período

c a las que se les ha aplicado la tarifa i , es decir, el número de pernотaciones por las que se recibió la tarifa i .

Las ponderaciones

$$\omega_i = \frac{P_t^0 N_i^c}{\sum_{i=1}^6 P_i^0 N_i^c}$$

se toman a partir de la misma encuesta. En ésta se les solicita a los hoteleros que indiquen el porcentaje de aplicación de cada una de las tarifas.

El IPH permite comparar la evolución de los precios facturados por las tarifas mencionadas con anterioridad, si bien no posibilita realizar comparaciones entre regiones o entre las diferentes tarifas en valores absolutos.

El IEA realiza la Encuesta de Coyuntura Turística de Andalucía (ECTA). Esta encuesta tiene como objetivo hacer frente a las necesidades y carencias de información relevantes en el sector del turismo en Andalucía.

Se trata de una encuesta de periodicidad trimestral y la consulta se dirige al conocimiento del perfil del turista que visita Andalucía, gasto que realiza y valoración del viaje.

19.5. Otras encuestas públicas

El Banco de España y el Departamento de Aduanas de la Agencia Tributaria realizan encuestas para hacer el seguimiento a la balanza comercial española.

El ministerio de Educación realiza operaciones estadísticas para conocer el estado de la Educación en los niveles de educación obligatoria, secundaria y universidad.

La Secretaría de Medio Ambiente que se encarga de evaluar a través de diversas operaciones estadísticas una serie de parámetros como el número de incendios, la calidad del aire, de las aguas, etc.

El Centro de Investigaciones Sociológicas (CIS), realiza encuestas sobre cuestiones de actualidad y valoración de la situación política, social y económica.

También la actividad de los juzgados y tribunales de los distintos órdenes judiciales, salvo lo social, es recogida en diversas Estadísticas Judiciales.

Diversos ministerios editan anuarios y boletines que recogen los datos más significativos de cada ministerio, como el Anuario Estadístico de Andalucía, el Anuario Estadístico de la Educación, el Anuario Estadístico de la Cultura, el Anuario Estadístico del Ministerio de Justicia, el Anuario de Estadísticas Laborales y de Asuntos Sociales, elaborados por los distintos ministerios y organismos del mismo nombre, el Boletín Estadístico Mensual del Banco de España, el Anuario de Estadísticas Agrarias y Pesqueras publicado por la Consejería de Agricultura y Pesca, que recopilan y difunden datos de sus respectivas competencias.

También son importantes los siguientes bancos de datos:

- (a) El *Directorio Central de Empresas (DIRCE)*, constituye una fuente de información para analizar altas, bajas y permanencias de empresas. Está organizado por el INE. El DIRCE trata de resumir en un directorio único todas las empresas españolas, siendo su objetivo básico hacer posible la realización de encuestas por muestreo dirigidas precisamente a las empresas. El DIRCE es el marco esencial para las encuestas económicas.
- (b) El *Banco de datos de Información Territorial BITMAT*, donde se desglosan datos socioeconómicos, financieros, jurídicos e institucionales de los municipios, entre otros.
- (c) El *Banco del Sistema de Información Municipal de Andalucía (SIMA)*, ofrece gran cantidad de información estadística multitemática y multiteritorial: datos de entorno físico, demográfico, económico y social de cualquier ámbito territorial de Andalucía: regional, provincial, municipal. De igual manera, incorpora información relativa a otras comunidades autónomas, y a países de la Unión Europea. Además, una vez elaborada la consulta, permite obtener gráficos y mapas a medida.

Capítulo 20

Muestreo en poblaciones normales

20.1. La distribución χ^2 de Pearson

Definición 20.1 La distribución χ^2 de Pearson con n grados de libertad, es la distribución de una variable continua X , cuya función de densidad es:

$$f(x) = \frac{e^{-\frac{x}{2}} x^{(n/2-1)}}{2^{n/2} \Gamma\left(\frac{n}{2}\right)}, \quad \text{para } x > 0, n > 0,$$

donde $\Gamma(p) = \int_0^\infty e^{-x} x^{p-1} dx$.

Simbólicamente escribiremos $X \rightsquigarrow \chi_n^2$.

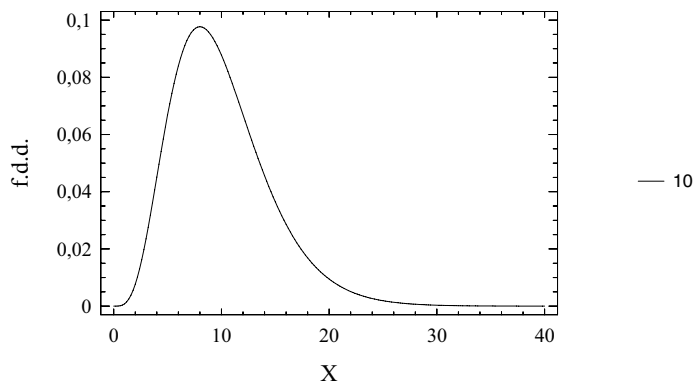
Teorema 20.1 Sean $X_1, X_2, \dots, X_n$, n variables aleatorias independientes entre sí, cada una con distribución $\mathcal{N}(0, 1)$.

Se verifica que

$$X_1^2 + X_2^2 + \dots + X_n^2 \rightsquigarrow \chi_n^2$$

Si $X \rightsquigarrow \chi_n^2$, sus principales características son:

- (a) $E[X] = n$.
- (b) $Var[X] = 2n$.
- (c) Sólo puede tomar valores positivos por tratarse de sumas de cuadrados de n variables aleatorias.

Figura 20.1: Función de densidad de una χ_{10}^2

Teorema 20.2 Sean $X \rightsquigarrow \chi_{n_1}^2$, $Y \rightsquigarrow \chi_{n_2}^2$ independientes entre sí. Se verifica que

$$X + Y \rightsquigarrow \chi_{n_1+n_2}^2.$$

(La variable es reproductiva respecto del parámetro n).

Teorema 20.3 Sea X una variable aleatoria con distribución χ_n^2 . Se verifica que si $n \rightarrow \infty$, la distribución de $\sqrt{2X} - \sqrt{2n-1}$ tiende a ser $\mathcal{N}(0,1)$.

(En la práctica se obtiene una aproximación aceptable si: $n > 30$).

En las tablas se encuentran las áreas a la derecha del punto crítico $\chi_{\alpha;n}^2$, es decir,

$$p[\chi_n^2 \geq \chi_{\alpha;n}^2] = \alpha$$

A α se le llama *nivel de significación*.

EJEMPLO 20.1 Dada una v.a. que sigue una distribución χ^2 de Pearson, podemos comprobar que:

(a) $\chi_{0,90;5}^2 = 1,610$

(b) $\chi_{0,01;26}^2 = 45,642$

(c) $p(\chi_8^2 \geq 3,49) = 0,90$

(d) $p(\chi_8^2 \leq 14) = 1 - p(\chi_8^2 > 14) = 1 - \alpha$, donde $\alpha = p(\chi_8^2 > 14)$ sale de interpolar

Abscisas	Áreas
$\chi_{0,10;8}^2 - \chi_{0,05;8}^2$	0,10 - 0,05
$\chi_{\alpha;8}^2 - \chi_{0,05;8}^2$	$\alpha - 0,05$

<i>Abcisas</i>	<i>Áreas</i>
13,362 – 15,507	0,10 – 0,05
14 – 15,507	α – 0,05

luego tras resolver la regla de tres se obtiene que

$$p(\chi_8^2 \leq 14) = 1 - p(\chi_8^2 > 14) = 1 - \alpha = 1 - 0,0851 = 0,9149$$

20.1.1. Distribución de la varianza muestral

Teorema 20.4 Sea X una v.a. $\mathcal{N}(\mu, \sigma)$, sea $(X_1, X_2, \dots, X_n)$ una m.a.s. de tamaño n , entonces

$$\frac{nS^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2$$

20.2. Distribución t de Student

Definición 20.2 Sea X_1 una variable aleatoria con distribución $\mathcal{N}(0, 1)$, y sea X_2 otra variable con distribución χ_n^2 , ambas independientes entre sí.

La variable definida como

$$X = \frac{X_1}{\sqrt{\frac{X_2}{n}}}$$

se dice que sigue una distribución t de Student con n grados de libertad.

Su función de densidad viene dada por:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in \mathbb{R}$$

Simbólicamente escribiremos $X \rightsquigarrow t_n$.

Sus principales características son:

- (a) $E[X] = 0$.
- (b) $Var[X] = \frac{n}{n-2}$, para $n > 2$.
- (c) El número de grados de libertad n es el número de variables aleatorias que figuran en el denominador.
- (d) Su función de densidad es simétrica.
- (e) Su campo de variabilidad es toda la recta real.

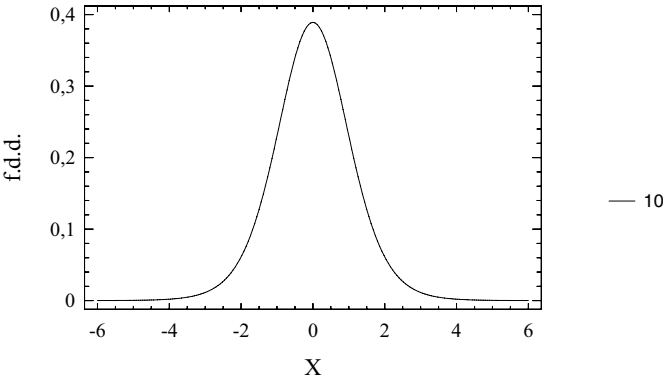


Figura 20.2: Función de densidad de una t_{10}

(f) Al aumentar el número de grados de libertad, la gráfica de la función se va haciendo más apuntada, siendo la distribución límite cuando $n \rightarrow \infty$ la $\mathcal{N}(0, 1)$.

En las tablas se encuentran las áreas a la derecha del punto crítico $t_{\alpha;n}$, es decir,

$$p[t \geq t_{\alpha;n}] = \alpha$$

A α se le llama *nivel de significación*.

EJEMPLO 20.2 Dada una v.a. que sigue una t de Student, se pide comprobar que:

(a) $t_{0,30;10}$, no se encuentra en las tablas, luego hay que interpolar

Abcisas	Áreas
$t_{0,40;10} - t_{0,25;10}$	0,40 – 0,25
$t_{0,30;10} - t_{0,25;10}$	0,30 – 0,25

Abcisas	Áreas
0,2602 – 0,6998	0,40 – 0,25
$t_{0,30;10} - 0,6998$	0,30 – 0,25

donde tras resolver la regla de tres se obtiene $t_{0,30;10} = 0,5532$.

(b) $p(t_8 \leq 1,2) = 1 - p(t_8 > 1,2) = 1 - \alpha = 1 - 0,134 = 0,866$, tras interpolar

Abcisas	Áreas
$t_{0,15;8} - t_{0,10;8}$	0,15 – 0,10
$t_{\alpha;8} - t_{0,10;8}$	$\alpha - 0,10$

20.2.1. Distribución del estadístico media muestral

(a) Si la varianza poblacional es conocida:

Sea X una v.a. $\mathcal{N}(\mu, \sigma)$ con σ conocida y sea $(X_1, X_2, \dots, X_n)$ una m.a.s. entonces el estadístico

$$\bar{X} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

y por tanto

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0, 1)$$

(b) Si la varianza es desconocida y la muestra es pequeña ($n < 30$):

Sea X una v.a. $\mathcal{N}(\mu, \sigma)$ con σ desconocida y sea $(X_1, X_2, \dots, X_n)$ una m.a.s. entonces como

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0, 1)$$

y

$$\frac{nS^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2$$

entonces por definición de la t de Student:

$$\frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{nS^2}{\sigma^2(n-1)}}} = \frac{\bar{X} - \mu}{S} \sqrt{n-1} \rightsquigarrow t_{n-1}$$

(c) Si la varianza es desconocida y la muestra es grande ($n \geq 30$):

Sea X una v.a. $\mathcal{N}(\mu, \sigma)$ con σ desconocida y sea $(X_1, X_2, \dots, X_n)$ una m.a.s. grande ($n \geq 30$)

$$\frac{\bar{X} - \mu}{S} \sqrt{n-1} = \frac{\bar{X} - \mu}{\hat{\sigma}} \sqrt{n} \rightsquigarrow \mathcal{N}(0, 1)$$

donde $\hat{\sigma} = S_c$ es el estadístico conocido como cuasi-desviación típica:

$$\hat{\sigma} = S_c = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}} = \frac{\sqrt{n}}{\sqrt{n-1}} S = \frac{\sqrt{n}}{\sqrt{n-1}} \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}}$$

20.2.2. Distribución de la diferencia de medias muestrales

(a) Si las varianzas poblacionales son conocidas:

Sean $(X_1, X_2, \dots, X_{n_1})$ y $(Y_1, Y_2, \dots, Y_{n_2})$ dos m.a.s tomadas de dos poblaciones independientes $\mathcal{N}(\mu_1, \sigma)$ y $\mathcal{N}(\mu_2, \sigma)$ se verifica:

$$\bar{X} - \bar{Y} \rightsquigarrow \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

(b) Si las varianzas poblacionales son desconocidas, pero iguales, y las muestras son pequeñas:

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \sqrt{(n_1 + n_2 - 2) \frac{n_1 n_2}{n_1 + n_2}} \rightsquigarrow t_{n_1 + n_2 - 2}$$

(c) Si las varianzas son desconocidas y distintas y las muestras son grandes:

$$\bar{X} - \bar{Y} \rightsquigarrow \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}\right)$$

20.3. Distribución F de Fisher-Snedecor

Definición 20.3 Sean X_1 y X_2 dos variables aleatorias independientes con distribuciones respectivas $\chi_{n_1}^2$ y $\chi_{n_2}^2$. La variable definida como

$$X = \frac{X_1/n_1}{X_2/n_2}$$

se dice que sigue una distribución $\mathcal{F}$ de Fisher-Snedecor con n_1 y n_2 grados de libertad.

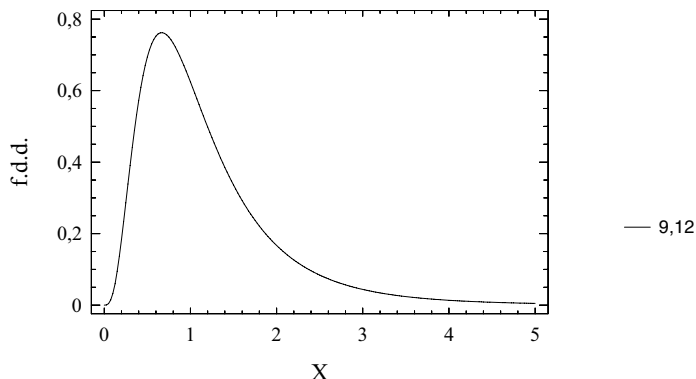
Su función de densidad viene dada por:

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right) \left(\frac{n_1}{n_2}\right)^{n_1/2}}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} \frac{x^{n_1/2-1}}{\left(1 + \frac{n_1 x}{n_2}\right)^{(n_1+n_2)/2}}, \quad x \geq 0, n_1 > 0, n_2 > 0.$$

Simbólicamente escribiremos $X \rightsquigarrow \mathcal{F}_{n_1, n_2}$.

Sus principales características son:

(a) $E[X] = \frac{n_2}{n_2 - 2}$, para $n_2 > 2$.

Figura 20.3: Función de densidad de una $\mathcal{F}_{9;12}$

$$(b) \text{Var}[X] = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}, \text{ para } n_2 > 4.$$

(c) El campo de variabilidad es $(0, \infty)$.

(d) La distribución límite de una $\mathcal{F}_{n_1, n_2}$ cuando $n_2 \rightarrow \infty$ es una $\chi_{n_1}^2$.

En las tablas se encuentran las áreas a la derecha del punto crítico $\mathcal{F}_{\alpha; n_1, n_2}$, es decir,

$$p[\mathcal{F}_{n_1, n_2} \geq \mathcal{F}_{\alpha; n_1, n_2}] = \alpha.$$

A α se le llama *nivel de significación*.

EJEMPLO 20.3 Dada una v.a. que sigue una $\mathcal{F}$ de Fisher-Snedecor, se pide comprobar que:

$$(a) \mathcal{F}_{0,10;9,12} = 2,21$$

$$(b) \mathcal{F}_{0,05;5,24} = 2,62$$

$$(c) \text{ Como } \mathcal{F}_{\alpha; n_1, n_2} = \frac{1}{\mathcal{F}_{1-\alpha; n_2, n_1}} \Rightarrow \mathcal{F}_{0,90;28,30} = \frac{1}{\mathcal{F}_{0,10;30,28}} = \frac{1}{1,63} = 0,61$$

(d) $p(\mathcal{F}_{10,20} \geq 2)$ interpolando queda

Abcisas	Áreas
2,35 – 1,94	0,05 – 0,10
2 – 1,94	α – 0,10

$$\text{Luego } p(\mathcal{F}_{10,20} \geq 2) = 0,092$$

20.3.1. Distribución del cociente de varianzas muestrales

Sean $(X_1, X_2, \dots, X_{n_1})$ y $(Y_1, Y_2, \dots, Y_{n_2})$ dos m.a.s tomadas de dos poblaciones independientes $\mathcal{N}(\mu_1, \sigma_1)$ y $\mathcal{N}(\mu_2, \sigma_2)$ se verifica:

$$\frac{n_1 S_1^2}{\sigma^2} \rightsquigarrow \chi_{n_1-1}^2$$

y

$$\frac{n_2 S_2^2}{\sigma^2} \rightsquigarrow \chi_{n_2-1}^2$$

Entonces por definición de la $\mathcal{F}$

$$\frac{\frac{n_1 S_1^2}{\sigma^2 (n_1 - 1)}}{\frac{n_2 S_2^2}{\sigma^2 (n_2 - 1)}} = \frac{n_1 (n_2 - 1) S_1^2}{n_2 (n_1 - 1) S_2^2} \rightsquigarrow \mathcal{F}_{n_1-1, n_2-1}$$

Capítulo 21

Estimación

21.1. Estimación puntual paramétrica

Consideremos una población descrita por la v.a. X cuya función de distribución $F(x, \theta)$ tiene forma conocida, desconociéndose únicamente el parámetro θ .

Definición 21.1 *Se llama estimación puntual al proceso de estimación que asigna a cada parámetro un cierto valor estimado.*

Un estimador puntual es una función de la muestra (estadístico), que ha sido construido con el objetivo de obtener un valor próximo(estimación), al parámetro desconocido de la población de la que se tomó la muestra. Un estimador del parámetro θ lo notaremos por $\hat{\theta} = \theta(X_1, \dots, X_n)$.

Hay distintas formas de conseguir estimadores:

21.1.1. El método analógico

Definición 21.2 *Este método consiste en tomar como estimador el análogo al correspondiente parámetro poblacional que se desea estimar.*

Parámetros	Estimadores
Media μ	$\hat{\mu} = \bar{X}$
Diferencia de medias $\mu_1 - \mu_2$	$\widehat{\mu_1 - \mu_2} = \bar{X}_1 - \bar{X}_2$
Varianza σ^2	$\hat{\sigma}^2 = S^2$
Desviación típica σ	$\hat{\sigma} = S$
Momentos α_k	$\hat{\alpha}_k = a_k = \frac{\sum X_i^k}{n}$
Proporción p	$\hat{p}$

OBSERVACIÓN 21.1 *Los estimadores son v.a. y los parámetros poblacionales son constantes.*

21.1.2. El método de los momentos

Este método consiste en igualar un cierto número de momentos muestrales con respecto al origen a los correspondientes de la población.

Si $a_r = \frac{\sum X_i^r}{n}$ es el momento muestral de orden r y $\alpha_r = E(X^r)$ es el momento poblacional de orden r , entonces el sistema de ecuaciones resultante es:

$$\left\{ \begin{array}{l} \bar{X} = \mu \\ \frac{\sum X_i^2}{n} = \alpha_2 \\ \vdots \end{array} \right.$$

Se obtiene así un sistema de ecuaciones donde figuran los parámetros que queremos estimar de incógnitas, y resolviendo el sistema obtenemos la expresiones deseadas.

EJEMPLO 21.1 *Obtégase un estimador del parámetro p de la distribución $\mathcal{B}(m, p)$, por el método de los momentos.*

Solución:

Obtenida una muestra de tamaño n , igualamos los momentos muestrales a los momentos poblacionales: $a_1 = \alpha_1 \Rightarrow \bar{X} = E(X) \Rightarrow \bar{X} = mp \Rightarrow \hat{p} = \frac{\bar{X}}{m}$.

EJEMPLO 21.2 *Obtégase un estimador de la varianza poblacional σ^2 por el método de los momentos de una población normal.*

Solución:

Obtenida una muestra de tamaño n , igualamos los momentos muestrales a los momentos poblacionales:

$$\left. \begin{array}{l} \bar{X} = \mu \\ \frac{\sum X_i^2}{n} = \alpha_2 = \sigma^2 + \mu^2 \end{array} \right\} \Rightarrow \hat{\sigma}^2 = \frac{\sum X_i^2}{n} - \bar{X}^2 = S^2.$$

Observemos que coincide con el estimador analógico correspondiente.

21.2. Estimación por intervalos de confianza

21.2.1. Concepto de intervalo de confianza

En secciones anteriores estudiábamos estimadores puntuales que nos permitían conocer los parámetros desconocidos de una determinada distribución.

Recordemos que los estimadores puntuales estiman los parámetros poblacionales puntualmente, es decir, son funciones de la muestra, que hacen corresponder a cada valor de la muestra un único número, de entre los posibles valores que puede tomar el parámetro.

No obstante, rara vez coincidirá esta estimación puntual con el verdadero valor del parámetro desconocido.

Por ejemplo, si queremos estimar la estatura media, θ , de los individuos de una determinada población, supuesto que ésta sigue una distribución normal $\mathcal{N}(\theta, 1)$. Como hemos visto, estimamos θ mediante $\bar{x}$ (la media de la muestra seleccionada), y rara vez ocurrirá verdaderamente que $\bar{x} = \theta$.

Es sin duda mucho más interesante, concluir la inferencia con un intervalo de posibles valores del parámetro (al que denominaremos intervalo de confianza), de forma que, el verdadero y desconocido valor del parámetro, se encuentre en dicho intervalo con una probabilidad todo lo alta que deseemos.

Así, por ejemplo, es mucho más deseable concluir afirmando que la media poblacional θ está entre $\bar{x} - 0,1$ y $\bar{x} + 0,1$, con probabilidad 0.99, que diciendo que la media muestral $\bar{x}$ es un buen estimador de θ .

Definición 21.3 *Se llama intervalo de confianza al intervalo dentro del cual esperamos quede el valor del parámetro θ , desconocido, con un grado de confianza medible, es decir, expresable numéricamente. Esta idea expresada más formalmente quedaría así:*

Sea $X_1, X_2, \dots, X_n$ una m.a.s. de una v.a. X cuya distribución depende de un parámetro θ . Se dice que los estadísticos

$$U = U(X_1, X_2, \dots, X_n) \qquad V = V(X_1, X_2, \dots, X_n)$$

constituyen un intervalo de confianza para θ , con nivel de confianza (o coeficiente de confianza) $1 - \alpha$, o al $(1 - \alpha)100\%$, si se verifica que:

$$p(U < \theta < V) = 1 - \alpha$$

Obviamente debe ser $U < V$ para toda la muestra de tamaño n .

OBSERVACIÓN 21.2

- (a) $1 - \alpha$ es la probabilidad de que el intervalo aleatorio (U, V) incluya al verdadero y desconocido valor del parámetro θ .
- (b) A α se le llama **nivel de significación** y se interpreta como la probabilidad de que el verdadero valor del parámetro no pertenezca al intervalo.
- (c) El nivel de confianza es un valor que suele tomar el experimentador que nos interesará lo más próximo a 1. Sin embargo en la práctica a mayor grado de confianza, menor precisión de la estimación (más grande es el intervalo). Por ejemplo, decir que se está 100 % seguro de que un pez está entre una orilla y otra del río, hace que esta ayuda sea tan vaga que resulta inútil.

EJEMPLO 21.3

<i>Pregunta del cliente</i>	<i>Respuesta del dependiente</i>	<i>Nivel implícito de confianza</i>	<i>Intervalo de confianza</i>
¿Recibiré mi lavadora en 1 año?	Estoy totalmente seguro de eso	Mayor que 99 %	1 año
¿Recibiré mi lavadora en 1 mes?	Estoy muy seguro de que se la llevamos este mes	Cerca del 95 %	1 mes
¿Recibiré mi lavadora en 1 semana?	Estoy casi seguro de montársela en 1 semana	Cerca del 80 %	1 semana
¿Recibiré mi lavadora mañana?	No estoy seguro de que la reciba entonces	Cerca del 40 %	1 día
¿Llegará mi lavadora a casa antes que yo?	Hay pocas probabilidades de que eso suceda	Cerca del 1 %	1 hora

- (d) U y V son estimadores por defecto y exceso de θ , que denominaremos límites de confianza de θ .
- (e) El valor de θ es constante, pero desconocido, mientras que el intervalo de confianza (U, V) es de extremos aleatorios.
- (f) Una vez obtenido el intervalo de confianza (a, b) ($U=a$ y $V=b$, al sustituir la m.a.s. por una realización muestral), éste ya no es de extremos aleatorios, no tiene sentido decir que cubre a θ con probabilidad $1 - \alpha$. La interpretación ha de ser **frecuencial**, por ejemplo, si $1 - \alpha = 0,95$, en una serie larga de determinaciones del intervalo (a, b) , cada vez con distintos valores muestrales, el $(1 - \alpha)100\% = 95\%$ de ellos cubrirán al verdadero valor de θ .
- (g) El principio de equivalencia estadística consiste en que cualquiera de los valores que pertenecen al intervalo pueden tomarse como estimación del parámetro.

21.2.2. Método del pivote

Es uno de los métodos más utilizados para construir intervalos de confianza. Consta de los siguientes pasos:

- (a) Buscamos un estadístico $T(X_1, X_2, \dots, X_n; \theta)$, que depende únicamente de la muestra y del parámetro desconocido, llamado *estadístico pivote*, del que conozcamos su distribución.

- (b) Entonces podemos calcular a y b tal que

$$p(a < T < b) = 1 - \alpha$$

- (c) Despejando θ (ya que T depende de θ), es posible encontrar dos funciones U y V tales que

$$p(U < \theta < V) = 1 - \alpha$$

21.3. Intervalo para la media

Consideremos una población descrita por la v.a. $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$ y sea $(X_1, X_2, \dots, X_n)$ una m.a.s. extraída de dicha población.

21.3.1. Varianza conocida

- (a) El estadístico pivote será

$$T(X_1, X_2, \dots, X_n; \theta) = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \rightsquigarrow \mathcal{N}(0, 1)$$

- (b) Construimos un intervalo para dicho estadístico con nivel de confianza $1 - \alpha$

$$p\left(a < \frac{\bar{X} - \mu}{\sigma} \sqrt{n} < b\right) = p\left(-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma} \sqrt{n} < z_{\alpha/2}\right) = 1 - \alpha$$

donde $z_{\alpha/2}$ es el punto crítico de una $\mathcal{N}(0, 1)$, tal que $p(Z \geq z_{\alpha/2}) = \alpha/2$.

- (c) Despejando el parámetro μ

$$p\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Luego el intervalo de confianza para la media

$$(U, V) = \left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right).$$

OBSERVACIÓN 21.3 La longitud del intervalo $L = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ depende de:

- (a) *Cuanto mayor sea la desviación típica poblacional mayor longitud tendrá el intervalo. Esto resulta convincente ya que, cuanto más dispersa esté la distribución de la población alrededor de su media, más incierta será nuestra inferencia sobre la media. Esta incertidumbre adicional se refleja en unos intervalos de confianza de mayor longitud.*

- (b) *Cuanto mayor sea el tamaño de la muestra menor será la amplitud del intervalo. Cuánta mayor información obtengamos sobre la población, más precisa será nuestra inferencia y eso se traduce en intervalos más cortos.*
- (c) *Cuanto mayor nivel de confianza mayor será la longitud del intervalo. Ya que a mayor nivel de confianza menor nivel de significación y por tanto mayor será $z_{\alpha/2}$.*

Cuando decimos que $\mu = \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ con un nivel de confianza $1-\alpha$ estamos admitiendo un error máximo de

$$\varepsilon = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Observamos que el error máximo admitido depende del nivel de significación y del tamaño de la muestra.

Fijados inicialmente un error máximo ε y un nivel de confianza, podemos deducir el tamaño de la muestra, simplemente despejando de la anterior fórmula

$$n = (z_{\alpha/2})^2 \frac{\sigma^2}{\varepsilon^2}.$$

EJEMPLO 21.4 *Los profesores de Estadística Administrativa de la Facultad de Ciencias Sociales y de la Comunicación de Jerez de la Frontera quieren determinar el tiempo medio que un alumno tarda en resolver un problema. Usando una m.a.s. de 36 alumnos registraron los siguientes datos:*

17	13	18	19	17	21	29	22	16	28	21	15
26	23	24	20	8	17	17	21	32	18	25	22
16	10	20	22	19	14	30	22	12	24	28	11

La media de la muestra es $\bar{x} = 19,9$ min. y de otras pruebas se sabe que la desviación típica poblacional es de 6 min.

- (a) *Dar una estimación puntual de la media poblacional μ .*
- (b) *Dar un intervalo de confianza al 95% de confianza para μ .*
- (c) *¿Cuál es el error máximo de la estimación?*
- (d) *Determinése el tamaño de la muestra necesario para estimar el tiempo medio de la población μ , si se quiere tener una confianza del 95% de que el error de estimación se sitúe como mucho entre ± 1 minuto.*

Solución:

- (a) *Una estimación puntual de $\mu = 19.9$ min.*

(b) Como σ es conocida, un intervalo de confianza al 95 %

$$p\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

sustituyendo

$$p\left(19,9 - z_{0,05/2} \frac{6}{\sqrt{36}} < \mu < 19,9 + z_{0,05/2} \frac{6}{\sqrt{36}}\right) = 1 - 0,05$$

como $z_{0,05/2} = z_{0,025} = 1,96$, entonces

$$p\left(19,9 - 1,96 \frac{6}{\sqrt{36}} < \mu < 19,9 + 1,96 \frac{6}{\sqrt{36}}\right) = 0,95$$

Luego el intervalo de confianza para la media al 95 %

$$\begin{aligned} 19,9 - 1,96 \frac{6}{\sqrt{36}} < \mu < 19,9 + 1,96 \frac{6}{\sqrt{36}} \\ 17,94 < \mu < 21,86 \end{aligned}$$

$$\boxed{I = (17,94; 21,86)}$$

Este intervalo puede contener a μ o no contenerla, pero si debiéramos apostar, serían 95 a 5 las posibilidades justas de que la contengan, es decir, este método funciona en un 95 % de las veces.

(c) El error máximo de la estimación es de $\varepsilon = 1,96 \frac{6}{\sqrt{36}} = 1,96$ al 95 % de confianza y para un tamaño muestral de 36.

(d) El error máximo de la estimación en este caso es de $\varepsilon = 1$ al 95 % de confianza. Como $1 - \alpha = 0,95 \Rightarrow z_{0,025} = 1,96$ y por tanto

$$n = (z_{\alpha/2})^2 \frac{\sigma^2}{\varepsilon^2} = (1,96)^2 \frac{6^2}{1^2} = 138,2$$

o sea, necesitamos 139 elementos muestrales.

21.3.2. Varianza desconocida y muestra pequeña

(a) El estadístico pivote será

$$T(X_1, X_2, \dots, X_n; \theta) = \frac{\bar{X} - \mu}{S} \sqrt{n-1} \rightsquigarrow t_{n-1}$$

(b) Construimos un intervalo para dicho estadístico con nivel de confianza $1-\alpha$,

$$p\left(a < \frac{\bar{X} - \mu}{S} \sqrt{n-1} < b\right) = \\ = p\left(-t_{\alpha/2; n-1} < \frac{\bar{X} - \mu}{S} \sqrt{n-1} < t_{\alpha/2; n-1}\right) = 1 - \alpha$$

donde $t_{\alpha/2; n-1}$ es el punto crítico de la distribución t de Student con $n-1$ grados de libertad, es decir $p(t_{n-1} \geq t_{\alpha/2; n-1}) = \alpha/2$.

(c) Despejando el parámetro μ

$$p\left(\bar{X} - t_{\alpha/2; n-1} \frac{S}{\sqrt{n-1}} < \mu < \bar{X} + t_{\alpha/2; n-1} \frac{S}{\sqrt{n-1}}\right) = 1 - \alpha$$

Luego el intervalo de confianza para la media

$$(U, V) = \left(\bar{X} - t_{\alpha/2; n-1} \frac{S}{\sqrt{n-1}}, \bar{X} + t_{\alpha/2; n-1} \frac{S}{\sqrt{n-1}}\right).$$

21.3.3. Varianza desconocida y muestra grande

$$(U, V) = \left(\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n-1}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n-1}}\right)$$

EJEMPLO 21.5 *En la oficina de control de pesas y medidas del mercado de una gran ciudad se presentó una denuncia relativa a que las botellas de "un litro" de cierta marca de leche embotellada no tenían la capacidad real de un litro. Los expendedores se defendieron alegando que era muy difícil que una botella en particular tuviese una capacidad de un litro, pero que la media de la población de botellas sí era de un litro.*

Para atender a la denuncia, se midieron cuidadosamente 100 botellas, elegidas al azar, y se calcularon la media y la desviación típica de sus capacidades, obteniendo los valores $\bar{x} = 995 \text{ cm}^3$ y $s = 10 \text{ cm}^3$. Suponiendo que la población de las capacidades de las botellas es normal, hallar unos límites de confianza de μ al nivel del 99 %. Interprete el resultado obtenido.

Solución:

El intervalo de confianza para la media de una población normal con desviación típica desconocida y muestra de tamaño grande,

$$(U, V) = \left(\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n-1}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n-1}}\right)$$

Como $1 - \alpha = 0,99 \Rightarrow \alpha = 0,01 \Rightarrow \alpha/2 = 0,005 \Rightarrow z_{\alpha/2} = z_{0,005} = 2,58$.

Luego el intervalo particular, no aleatorio, es:

$$\left(995 - 2,58 \frac{10}{\sqrt{100-1}}, 995 + 2,58 \frac{10}{\sqrt{100-1}} \right) = (992,407; 997,5529).$$

Como 1 litro=1000 cm³ no pertenece al intervalo, se puede concluir que la media de las capacidades de todas la botellas no es igual a un litro y por tanto el denunciante tenía razón (con un 99 % de confianza).

21.4. Intervalo de confianza para la varianza

(a) Con media desconocida el estadístico pivote será

$$\frac{nS^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2$$

(b) Construimos un intervalo para dicho estadístico con nivel de confianza $1-\alpha$

$$p\left(a < \frac{nS^2}{\sigma^2} < b\right) = p\left(\chi_{1-\alpha/2}^2 < \frac{nS^2}{\sigma^2} < \chi_{\alpha/2}^2\right) = 1 - \alpha$$

(c) Despejando el parámetro σ^2

$$p\left(\frac{\chi_{1-\alpha/2}^2}{nS^2} < \frac{1}{\sigma^2} < \frac{\chi_{\alpha/2}^2}{nS^2}\right) = p\left(\frac{nS^2}{\chi_{\alpha/2}^2} < \sigma^2 < \frac{nS^2}{\chi_{1-\alpha/2}^2}\right) = 1 - \alpha$$

Luego el intervalo de confianza para la varianza

$$(U, V) = \left(\frac{nS^2}{\chi_{\alpha/2; n-1}^2}, \frac{nS^2}{\chi_{1-\alpha/2; n-1}^2} \right)$$

EJEMPLO 21.6 Como supervisor de un proceso de empaquetado de sobres de azúcar de la AZUCARERA DE JEREZ S.A., se han tomado 12 sobres cuya media es de $\bar{x} = 15,97$ gramos y varianza $s^2 = 0,0224$. Utilizando intervalos de confianza del 90 % calcule para todos los sobres de azúcar que se empaquetan en la planta:

(a) Un intervalo para la varianza.

(b) Idem para la desviación típica. ¿Puede ser la desviación típica igual a uno?

Solución:

(a) Luego el intervalo de confianza para la varianza

$$(U, V) = \left(\frac{nS^2}{\chi_{\alpha/2; n-1}^2}, \frac{nS^2}{\chi_{1-\alpha/2; n-1}^2} \right)$$

Como $\chi_{\alpha/2; n-1}^2 = \chi_{0,05; 11}^2 = 19,675$; $\chi_{1-\alpha/2; n-1}^2 = \chi_{0,95; 11}^2 = 4,575$

Luego,

$$I = \left(\frac{12(0,0224)}{19,675}, \frac{12(0,0224)}{4,575} \right) = (0,0136; 0,0587)$$

(b) Luego el intervalo de confianza para la desviación típica

$$(U, V) = \left(\sqrt{\frac{nS^2}{\chi_{\alpha/2; n-1}^2}}, \sqrt{\frac{nS^2}{\chi_{1-\alpha/2; n-1}^2}} \right)$$

Luego,

$$I = \left(\sqrt{\frac{12(0,0224)}{19,675}}, \sqrt{\frac{12(0,0224)}{4,575}} \right) = (0,1166; 0,2422)$$

Como $1 \notin I$ entonces aceptaremos que la desviación típica es distinta de uno al 90 % de confianza.

21.5. Intervalo para las medias

21.5.1. Varianzas conocidas

(a) Consideremos dos poblaciones normales independientes, luego

$$\bar{X}_1 - \bar{X}_2 \rightsquigarrow \mathcal{N} \left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

entonces el estadístico pivote será:

$$T(X_1, X_2, \dots, X_n; \theta) = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0, 1)$$

(b) Construimos un intervalo para dicho estadístico con nivel de confianza $1-\alpha$,

$$p \left(a < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < b \right) = 1 - \alpha$$

$$p \left(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{\alpha/2} \right) = 1 - \alpha$$

(c) Despejando $\mu_1 - \mu_2$

$$(U, V) = \left((\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

21.5.2. Varianzas desconocidas, iguales y muestras pequeñas

$$(U, V) = \left((\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2; n_1+n_2-2} \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 n_2}} \right)$$

21.5.3. Varianzas desconocidas, distintas y muestras pequeñas

$$(U, V) = \left((\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2; m} \sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}} \right)$$

siendo

$$c = \frac{\frac{S_1^2}{n_1 - 1}}{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}} \quad \frac{1}{m} = \frac{c^2}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1}$$

21.5.4. Varianzas desconocidas y muestras grandes

$$(U, V) = \left((\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}} \right)$$

EJEMPLO 21.7 *El número diario de piezas fabricado por una máquina A en cinco días ha sido 50, 48, 53, 60 y 37; mientras que, en esos mismos días, una máquina B fabricó respectivamente 40, 51, 62, 55 y 64. Supuesto que ambas poblaciones siguen aproximadamente una distribución normal, se pide:*

- Construye un intervalo para la diferencia de medias entre las máquinas A y B, suponiendo la misma desviación típica y nivel de confianza del 95 %.*
- Idem, pero sin suponer la misma desviación típica. ¿Puedo considerar las dos medias poblacionales iguales?*
- Determina cuál debía haber sido el tamaño muestral n común a ambas muestras para que, siendo las desviaciones típicas poblacionales $\sigma_A = 8$, $\sigma_B = 9,05$, la longitud del intervalo para la diferencia de medias con $\alpha = 0,05$ fuese de 8 unidades.*

Solución:

- Si las varianzas son desconocidas pero iguales y muestras pequeñas,*

$$(U, V) = \left((\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2; n_1+n_2-2} \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 n_2}} \right)$$

Calculemos primero con la información que nos dan las muestras

$$50, 48, 53, 60, 37 \Rightarrow \bar{x}_1 = 49,6; s_1^2 = 56,24$$

$$40, 51, 62, 55, 64 \Rightarrow \bar{x}_2 = 54,4; s_2^2 = 73,84$$

$$\text{Como } 1-\alpha = 0,95 \Rightarrow \alpha/2 = 0,025 \Rightarrow t_{\alpha/2; n_1+n_2-2} = t_{0,025; 8} = 2,306.$$

Luego,

$$I = \left((49,6 - 54,4) \pm 2,306 \sqrt{\frac{5(56,24) + 5(73,84)}{5 + 5 - 2} \cdot \frac{5 + 5}{5 \cdot 5}} \right) = (-17,9; 8,3)$$

- Si las varianzas son desconocidas pero distintas y muestras pequeñas,*

$$(U, V) = \left((\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2; m} \sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}} \right)$$

siendo

$$c = \frac{\frac{S_1^2}{n_1 - 1}}{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}} = \frac{56,24}{56,24 + 73,84} = 0,4323$$

$$\frac{1}{m} = \frac{c^2}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1} \Rightarrow m = \frac{4}{2c^2 - 2c + 1} = 7,85 \simeq 8$$

por tanto $t_{\alpha/2;m} = t_{0,025;8} = 2,306$.

$$I = \left((49,6 - 54,4) \pm 2,306 \sqrt{\frac{56,24}{5-1} + \frac{73,84}{5-1}} \right) = (-17,9; 8,3).$$

En este caso particular, este intervalo coincide con el anterior, pero no tiene siempre que ser así.

Como $0 \in I$ entonces puedo considerar las dos medias poblacionales iguales.

(c) Dado $I=(a,b) \Rightarrow L = b - a$ donde las varianzas son conocidas

$$(U, V) = \left((\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

luego,

$$L = 2z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \Rightarrow 8 = 2(1,96) \sqrt{\frac{64}{n} + \frac{81,9025}{n}} \Rightarrow n = 35,0311 \simeq 35.$$

21.6. Intervalo para las medias de datos apareados

Sean $(X_1, X_2, \dots, X_n)$ e $(Y_1, Y_2, \dots, Y_n)$ son dos m.a.s. tomadas de dos poblaciones no independientes $\mathcal{N}(\mu_1, \sigma_1)$ y $\mathcal{N}(\mu_2, \sigma_2)$, un intervalo de confianza para la diferencias de medias en el caso de muestras pequeñas:

$$(U, V) = \left(\bar{d} \pm t_{\alpha/2;n-1} \cdot \frac{S_d}{\sqrt{n-1}} \right)$$

$$\text{donde } \bar{d} = \frac{\sum d_i}{n} = \frac{\sum (X_i - Y_i)}{n} \quad \text{y} \quad S_d^2 = \frac{\sum (d_i - \bar{d})^2}{n} = \frac{\sum d_i^2}{n} - (\bar{d})^2.$$

EJEMPLO 21.8 *Un equipo de investigación biológica está interesado en ver si una nueva droga reduce significativamente el nivel de colesterol en la sangre. Con tal fin, forma una muestra de 10 pacientes y determina el contenido en colesterol en la sangre antes y después del tratamiento. Los datos muestrales expresados en miligramos por 10 mililitros son los siguientes:*

Paciente	1	2	3	4	5	6	7	8	9	10
Antes	217	252	229	200	209	213	215	260	232	216
Después	209	241	230	208	206	211	209	228	224	203

Se pide, construir un intervalo de confianza del 95 % para la diferencia de contenido medio de colesterol en la sangre antes y después del tratamiento.

Solución:

Se trata de datos apareados, en los que no existe independencia entre las muestras

$$(U, V) = \left(\bar{d} \pm t_{\alpha/2; n-1} \cdot \frac{S_d}{\sqrt{n-1}} \right)$$

Con esos datos

$d_i = X_i - Y_i$	8	11	-1	-8	3	2	6	32	8	13
-------------------	---	----	----	----	---	---	---	----	---	----

luego

$$\bar{d} = 7,4; \quad s_d^2 = 100,84; \quad s_d = 10,04$$

$$n = 10; \quad \alpha/2 = 0,025 \Rightarrow t_{0,025;9} = 2,262$$

entonces

$$I = \left(7,4 \pm 2,262 \frac{10,04}{\sqrt{9}} \right) = (-0,1701; 14,97).$$

21.7. Intervalo para la razón de varianzas

$$(U, V) = \left(\frac{n_1(n_2-1)S_1^2}{n_2(n_1-1)S_2^2} \cdot \frac{1}{\mathcal{F}_{\alpha/2; n_1-1, n_2-1}}, \frac{n_1(n_2-1)S_1^2}{n_2(n_1-1)S_2^2} \cdot \mathcal{F}_{\alpha/2; n_2-1, n_1-1} \right)$$

EJEMPLO 21.9 El número diario de piezas fabricadas por una máquina A en cinco días tiene de media 50 y desviación típica 8; mientras que, en esos mismos días, una máquina B fabricó: 40, 51, 62, 55 y 64 piezas respectivamente. Suponiendo que ambas poblaciones siguen aproximadamente una distribución normal, se pide:

Construye un intervalo de estimación para el cociente de varianzas poblacionales con nivel de confianza de 0.95. ¿Podemos considerar las dos varianzas poblacionales iguales?

Solución:

Como el intervalo para el cociente de varianzas es

$$(U, V) = \left(\frac{n_1 (n_2 - 1) S_1^2}{n_2 (n_1 - 1) S_2^2} \cdot \frac{1}{\mathcal{F}_{\alpha/2; n_1-1, n_2-1}}, \frac{n_1 (n_2 - 1) S_1^2}{n_2 (n_1 - 1) S_2^2} \cdot \mathcal{F}_{\alpha/2; n_2-1, n_1-1} \right)$$

Necesitamos:

$$\bar{x}_1 = 50, \quad s_1 = 8, \quad n_1 = 5$$

$$\bar{x}_2 = \frac{272}{5} = 54,4; \quad s_2 = \sqrt{\frac{15166}{5} - (54,4)^2} = 8,5930; \quad n_2 = 5$$

$$1 - \alpha = 0,95 \Rightarrow \alpha/2 = 0,025 \Rightarrow \mathcal{F}_{0,025;4,4} = 9,6045$$

Luego

$$I = \left(\frac{(5)(4)(64)}{(5)(4)(73,84)} \cdot \frac{1}{9,6045}, \frac{(5)(4)(64)}{(5)(4)(73,84)} 9,6045 \right) = (0,0902; 8,3245)$$

Como $1 \in I$ entonces aceptaremos que las dos varianzas poblacionales son iguales al 95 % de confianza.

21.8. Intervalos de confianza asintóticos

Los siguientes intervalos son asintóticos en el sentido de que la distribución del estadístico la obtendremos pasando al límite y por tanto necesitaremos tamaños de muestras suficientemente grandes.

21.8.1. Intervalo de confianza para la proporción

En una población binomial, es decir $X \rightsquigarrow \mathcal{B}(1, p)$. Consideramos una m.a.s. de tamaño n .

Sabemos que

$$\hat{p} = \frac{\text{n}^\circ \text{ de éxitos}}{n} = \frac{\sum X_i}{n}$$

(a)

$$\hat{p} \rightsquigarrow \mathcal{N} \left(p, \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \text{ cuando } n \rightarrow \infty$$

El estadístico pivote será:

$$T(X_1, X_2, \dots, X_n; \theta) = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \rightsquigarrow \mathcal{N}(0, 1)$$

(b) Construimos un intervalo para dicho estadístico a un nivel de confianza $1-\alpha$,

$$p \left(a < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < b \right) = p \left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z_{\alpha/2} \right) = 1 - \alpha$$

(c) Despejando p

$$(U, V) = \left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Si se usa $\hat{p}$ como una estimación de p , podemos afirmar con una confianza del $1-\alpha$ que el error máximo es

$$\varepsilon = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Despejando n de la anterior expresión podemos dar el tamaño de la muestra para estimar p ,

$$n = \hat{p}(1-\hat{p}) \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2.$$

OBSERVACIÓN 21.4 Si se desconoce cualquier información sobre p entonces $\hat{p} = 0,5$.

EJEMPLO 21.10 Una nueva droga ha curado a 70 de 200 enfermos. Se pide:

- Estima, mediante un intervalo de confianza del 99 %, la proporción de personas curadas, si la nueva droga se hubiese aplicado a una población constituida por todos los individuos con la misma enfermedad.
- ¿Cuál es la longitud del intervalo de confianza al 95 %?
- ¿Cuál debe ser el tamaño muestral para estimar la proporción de personas curadas con una probabilidad de por lo menos 0.95 que su error no será mayor de 0.04, supuesto una estimación de p de 0.35?
- Contesta a la anterior pregunta si además se desconoce cuál puede ser una estimación de p . Compara el resultado con el del apartado último.

Solución:

- La población en estudio es dicotómica, es decir sus elementos o poseen una determinada propiedad o no la poseen, en este caso las personas enfermas o se curan o no se curan.

Construiremos un intervalo de confianza para p , proporción de personas enfermas curadas. Como $n = 200 \geq 30$ podemos hacer uso de un test asintótico

$$(U, V) = \left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

$$1-\alpha = 0,99 \Rightarrow \alpha/2 = 0,005 \Rightarrow z_{0,005} = 2,58 \quad y \quad \hat{p} = \frac{70}{200} = 0,35.$$

Entonces

$$I = \left(0,35 \pm 2,58 \sqrt{\frac{0,35(1-0,35)}{200}} \right) = (0,263; 0,437).$$

Luego la proporción de personas curadas en toda la población con un 99 % de confianza oscilará entre 26.3 % y 43.7 %.

(b) Utilizamos el mismo intervalo, pero variando el nivel de confianza,

$$(U, V) = \left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

$$1-\alpha = 0,95 \Rightarrow \alpha/2 = 0,025 \Rightarrow z_{0,025} = 1,96, \text{ entonces}$$

$$I = \left(0,35 \pm 1,96 \sqrt{\frac{0,35(1-0,35)}{200}} \right) = (0,284; 0,416).$$

Por tanto

$$L = 0,416 - 0,284 = 0,132.$$

Observamos que la longitud del intervalo ha disminuido al disminuir el nivel de confianza.

(c) El tamaño de la muestra para estimar p , conocido que $\hat{p} = 0,35$ para $1-\alpha = 0,95$ y $\varepsilon = 0,04$ es de

$$n = \hat{p}(1-\hat{p}) \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 = 0,35 \cdot 0,65 \cdot \left(\frac{1,96}{0,04} \right)^2 = 546,2 \Rightarrow n \simeq 547.$$

(d) El tamaño de la muestra para estimar p , desconocida cualquier información sobre p para $1-\alpha = 0,95$ y $\varepsilon = 0,04$ es de

$$n = 0,5(1-0,5) \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 = 0,5 \cdot 0,5 \cdot \left(\frac{1,96}{0,04} \right)^2 = 600,25 \Rightarrow n \simeq 601.$$

Observamos como cualquier conocimiento sobre p reduce sustancialmente el tamaño de la muestra requerido para lograr un grado de precisión deseado.

21.8.2. Intervalo para la diferencia de proporciones

Supongamos que un experimento dicotómico (2 resultados posibles), se realiza en dos poblaciones diferentes.

Sean p_1 y p_2 las proporciones poblacionales correspondientes a cada una de las poblaciones.

Supongamos que realizamos el experimento n_1 veces en la primera población y n_2 veces en la segunda.

Sabemos que:

$$\hat{p}_1 = \frac{n^0 \text{ de éxitos en la primera muestra}}{n_1}$$

$$\hat{p}_2 = \frac{n^0 \text{ de éxitos en la segunda muestra}}{n_2}$$

(a) Como

$$\hat{p}_1 - \hat{p}_2 \rightsquigarrow \mathcal{N} \left(p_1 - p_2, \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}} \right)$$

el estadístico pivote será:

$$T(X_1, X_2, \dots, X_n; \theta) = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}}} \rightsquigarrow \mathcal{N}(0, 1)$$

cuando $n_1 \rightarrow \infty$ y $n_2 \rightarrow \infty$.

(b) Construimos un intervalo para dicho estadístico con nivel de confianza $1-\alpha$,

$$p \left(a < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}}} < b \right) = 1 - \alpha$$

$$p \left(-z_{\alpha/2} < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}}} < z_{\alpha/2} \right) = 1 - \alpha.$$

(c) Despejando $p_1 - p_2$

$$(U, V) = \left((\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}} \right)$$

EJEMPLO 21.11 *Al intentar medir la opinión de los padres de un sistema escolar con respecto a un nuevo plan de estudios, un inspector escolar recopila muestras aleatorias de 100 padres de familia en cada una de las dos regiones residenciales incluidas en el sistema escolar. En la primera región 70 de los 100 padres expresaron su opinión de votar a favor de la proposición, mientras que en la segunda región, 50 de los 100 votantes muestreados indicaron su intención de hacerlo. Estima la diferencia entre las proporciones reales de padres de familia de las dos áreas que tienen intención de votar a favor de la proposición, utilizando límites de confianza del 95 %. ¿Podríamos considerar las dos proporciones poblacionales iguales?*

Solución:

La población es dicotómica: o votan a favor o no.

El intervalo que necesito es

$$(U, V) = \left((\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}} \right)$$

con $\hat{p}_1 = 0,7$, $\hat{p}_2 = 0,5$, $n_1 = n_2 = 100$, $1 - \alpha = 0,95 \Rightarrow z_{\alpha/2} = z_{0,025} = 1,96$, entonces

$$I = \left((0,7 - 0,5) \pm 1,96 \sqrt{\frac{0,7 (1 - 0,7)}{100} + \frac{0,5 (1 - 0,5)}{100}} \right) = (0,07; 0,33).$$

La diferencia en las proporciones poblacionales se encuentra entre el 7 y el 33 %.

Como $0 \notin I$ entonces considero las dos proporciones poblacionales distintas.

Capítulo 22

Contrastes de hipótesis

22.1. Introducción

Supongamos una cierta población, en la que se estudia una variable que se considera aleatoria, cuya función de densidad $f(X; \theta)$ depende de un cierto parámetro θ . Este parámetro toma valores dentro de un cierto espacio paramétrico Θ , de tal manera que para cada valor que toma θ en Θ , la función $f(X; \theta)$ es distinta.

Definición 22.1 *Una hipótesis estadística es una afirmación respecto a una característica de la población. Mas formalmente, una hipótesis estadística sobre el parámetro θ es una conjetura sobre el valor concreto que tenga dicho parámetro en la población.*

Por ejemplo, la afirmación: "*El tabaco produce cáncer*", es una hipótesis que se refiere al conjunto de elementos de una o varias poblaciones, en este caso, la población de fumadores y de no fumadores. La variable que estudiamos es el número de casos de cáncer entre los fumadores y no fumadores y lo que queremos es comparar proporción de casos de cáncer en esas dos poblaciones.

El establecimiento de una hipótesis sobre θ , supone dividir el espacio paramétrico Θ en dos partes, una denominada Θ_0 , integrada por el conjunto de valores que cumplen la hipótesis, y otra Θ_1 , integrada por el conjunto de valores que no la cumplen.

Definición 22.2 *Generalmente se marca la preferencia de ambas hipótesis, en el sentido que se considera a una de ellas como la más adecuada que la otra, y que mantendremos como cierta a no ser que se obtenga una evidencia de lo contrario y se le llama hipótesis nula H_0 ; y a la segunda se la conoce como hipótesis alternativa H_1 , y es sencillamente la negación de la hipótesis nula.*

$$H_0 : \theta \in \Theta_0$$

$$H_1 : \theta \in \Theta_1$$

El nombre de "nula" proviene de que H_0 representa la hipótesis que mantendremos a no ser que los datos indiquen su falsedad. "Nula" debe entenderse en el sentido de "neutra". No sabremos si es verdadera o no, a no ser que estudiemos todos los elementos de la población.

Definición 22.3 Si los subconjuntos Θ_0 ó Θ_1 están compuestos por un sólo elemento, la hipótesis es simple. Son el tipo de hipótesis que especifican un sólo valor para el parámetro ($\theta = \theta_0$).

Definición 22.4 Las hipótesis compuestas al contrario de las hipótesis simples, especifican un intervalo de valores para el parámetro.

De las dos posibles hipótesis alternativas compuestas, consideraremos dos tipos de contrastes:

- (a) *Contrastes bilaterales o de dos colas, si desconocemos en qué dirección puede ser falsa la hipótesis nula, es decir, la hipótesis alternativa tiene la forma $\theta \neq \theta_0$.*
- (b) *Contrastes unilaterales o de una cola, si conocemos la dirección en que la hipótesis nula puede ser falsa, en este caso, la forma que adopta la hipótesis alternativa es de la forma $\theta \geq \theta_0$ ($\theta \leq \theta_0$).*

Definición 22.5 Contraste o test de hipótesis es una regla de decisión mediante la cual optamos por una hipótesis u otra. Los contrastes son procedimientos que nos permiten determinar si las muestras observadas difieren significativamente de los resultados esperados.

Si suponemos que una hipótesis es cierta, pero vemos que los resultados obtenidos en la muestra difieren notablemente de los esperados bajo tal hipótesis, nos veremos obligados a rechazar la hipótesis de partida.

La metodología actual de contraste de hipótesis es el resultado de los trabajos de R.A. Fisher, J. Neymann y Egon Pearson entre 1920 y 1933. Su lógica es similar a la de un juicio penal, donde debe decidirse si el acusado es culpable o inocente. En un juicio, la hipótesis nula, es la que trataremos de mantener a no ser que los datos me demuestren lo contrario, es que el acusado es inocente. El juicio consiste en aportar evidencia suficiente para rechazar la hipótesis nula de inocencia más allá de cualquier duda razonable.

Análogamente, en un contraste de hipótesis se analiza si los datos muestrales permiten rechazar la hipótesis nula, es decir, si los datos observados tienen una probabilidad de aparecer lo suficientemente pequeña cuando la hipótesis nula es cierta.

Si la hipótesis nula especifica el parámetro de la distribución de una variable en una población, el contraste consiste en tomar una muestra aleatoria simple y calcular un estimador del parámetro. Si el estimador está "próximo" al valor del parámetro indicado por la hipótesis nula, concluiremos que la hipótesis ha predicho lo observado y que no existe evidencia para rechazarla.

Si, por el contrario, la diferencia entre ambas es grande, concluiremos que existe una discrepancia significativa entre lo previsto por la hipótesis y lo observado, y rechazaremos la hipótesis nula.

Se ha generalizado el uso del verbo "*aceptar*" en lugar de "*no rechazar*". Significa sencillamente que, cuando los datos de la muestra no nos llevan a rechazar la hipótesis nula, nos portamos como si ésta fuera verdadera. La única manera de "*aceptarla*" definitivamente sería efectuar un censo de toda la población.

Definición 22.6 *Llamaremos región crítica o de rechazo R_C para el contraste de una hipótesis, al conjunto de valores del estadístico de contraste que nos lleva a tomar la decisión de rechazar la hipótesis nula.*

Definición 22.7 *Definiremos región de aceptación R_A al conjunto de valores del estadístico de contraste que nos lleva a aceptar la hipótesis nula.*

OBSERVACIÓN 22.1

- (a) *La región de aceptación del contraste es mutuamente excluyente con la región de rechazo.*
- (b) *El rechazo de la hipótesis nula equivale a la aceptación de la alternativa, y viceversa.*
- (c) *Cuando se habla de aceptación o rechazo de la hipótesis nula, se debe entender en el sentido de que la muestra ha proporcionado una evidencia suficiente, nunca absoluta.*

	H_0 cierta	H_0 falsa
H_0 rechazar	Error tipo I= ε_1	Decisión correcta
H_0 aceptar	Decisión correcta	Error tipo II= ε_2

Definición 22.8 *Error tipo I: Error que se comete cuando se rechaza H_0 siendo cierta.*

Definición 22.9 *Error tipo II: Error que se comete cuando se acepta H_0 siendo falsa.*

Definición 22.10 *Nivel de significación o tamaño de la región crítica para el contraste*

$$\alpha = p(\varepsilon_1) = p(\text{rechazar } H_0 / H_0 \text{ es cierta}) = p(d_M \in R_C / H_0 \text{ es cierta})$$

donde d_M es el que llamaremos estadístico de discrepancia evaluado en la muestra. El nivel de significación es el punto crítico de probabilidad a partir del cual ya no estamos dispuestos a sostener que nuestro resultado muestral se derive del azar.

El nivel de significación se especifica antes de tomar la muestra, de manera que los resultados obtenidos no influyan en su elección.

Es posible probar cualquier hipótesis nula a cualquier nivel de significación. Cuánto más alto sea el nivel de significación que utilicemos al probar una hipótesis, mayores probabilidades habrá de rechazar una hipótesis nula aunque sea verdadera. Un nivel de significación de 0,50 es tan alto que rara vez aceptaremos una hipótesis cuando no sea verdadera ya que nuestra región de aceptación es muy pequeña; pero al mismo tiempo, esa seguridad tiene un precio: frecuentemente la rechazaremos aunque sea verdadera.

Y ante la imposibilidad de minimizar simultáneamente los dos errores (salvo que aumentemos el tamaño de la muestra, cosa que no siempre es posible), se opta por acotar un tipo de error y minimizar el otro.

Fijado el nivel de significación α , de todas las regiones críticas con el mismo nivel de significación, se elige aquella donde la potencia del contraste sea mayor, donde

Definición 22.11 *Definimos la potencia de un test como*

$$\text{Potencia del test} = 1 - \beta = 1 - p(\varepsilon_2) = p(\text{rechazar } H_0 / H_0 \text{ es falsa})$$

22.2. Pasos para la realización de un contraste

- (a) Determinar la hipótesis nula H_0 y la hipótesis alternativa H_1 .
- (b) Mediante una m.a.s. de tamaño n se recoge la información de la población.
- (c) Seleccionar una medida de discrepancia o estadístico de contraste $d = d(\theta_0, \hat{\theta})$, que mide la discrepancia entre el valor del parámetro propugnado por la hipótesis nula, θ_0 , y el que podríamos asignar basándonos en la muestra, $\hat{\theta}$. Para que la medida de discrepancia no dependa de las unidades de medida de la variable, se suele dividir esa diferencia por el valor promedio, que es el error típico de estimación del parámetro. Para ello necesitamos conocer la distribución en el muestreo del estadístico d si H_0 es verdadera.
- (d) Elegir el nivel de significación α .
- (e) Determinar la región crítica R_C determinada por d_C , siendo d_C el valor crítico del estadístico de contraste d bajo el supuesto de que la hipótesis nula es cierta. Este valor crítico, es el valor a partir del cual la diferencia entre θ_0 y $\hat{\theta}$ es demasiado grande para ser atribuido al azar

$$p(d \geq d_C / H_0) = \alpha$$

- (f) Cálculo del valor del estadístico $\hat{\theta}$ y de la medida de discrepancia d para una muestra particular, d_M .

- (g) Obtenemos conclusiones de tipo estadístico. Comparamos el valor observado del estadístico en la muestra, d_M , con el valor o valores críticos d_C .

Si d_M pertenece a la región de rechazo del contraste, esto es, si la diferencia entre lo que esperamos de acuerdo a la hipótesis nula y lo que observamos en la muestra es muy grande para ser atribuido al azar, entonces rechazaremos H_0 , es decir, la discrepancia es *significativa estadísticamente*.

Si por el contrario, d_M está en la región de aceptación del contraste, la diferencia entre lo que esperamos de acuerdo con la hipótesis nula y lo que observamos en la muestra es muy pequeña, decimos que el resultado *no es significativo estadísticamente*, es decir, aceptamos H_0 .

- (h) Obtenemos conclusiones de naturaleza no estadística.

OBSERVACIÓN 22.2

- (a) El término "diferencia significativa" tiene poco que ver con la significatividad práctica. Por ejemplo, si se trata de contrastar si la producción media de una máquina es μ_0 ; si tomamos una muestra muy grande, es bastante probable que observemos una diferencia significativa y rechazemos que la media es μ_0 . Sin embargo, la conclusión puede ser que la media es $\mu = \mu_0 + 0,00001$, y la diferencia entre μ y μ_0 puede ser irrelevante en la práctica. Para distinguir entre significatividad estadística y práctica conviene estimar el parámetro a la vista de los datos.
- (b) El resultado de un contraste puede depender mucho del valor de α , que es arbitrario, siendo posible rechazar H_0 con $\alpha = 0,05$ y aceptarlo con $\alpha = 0,04$.
- (c) Dar el resultado del contraste no permite diferenciar el grado de evidencia que la muestra indica a favor o en contra de H_0 . Se puede rechazar la hipótesis nula pero con una evidencia muy distinta.

Para hacer frente a las dos últimas observaciones deberemos utilizar en vez del nivel de significación el llamado *p*-valor.

Definición 22.12 Se define el *p*-valor del contraste como la probabilidad de obtener una discrepancia d , entre θ_0 y $\hat{\theta}$, mayor o igual que la observada en la muestra d_M , cuando H_0 es cierta,

$$p = p(d \geq d_M / H_0)$$

El *p*-valor es el nivel de significación más bajo que hace que la muestra observada nos indique que se podría rechazar la H_0 .

OBSERVACIÓN 22.3

- (a) Mientras que el nivel de significación se fija a priori, a partir de la opinión que tengamos de la validez de la hipótesis nula y de las consecuencias que tengamos de equivocarnos, el p-valor se determina a partir de la muestra, dejando que el investigador elabore sus propias consecuencias.
- (b) Un p-valor grande nos indica que nuestro resultado muestral no es muy diferente del resultado predicho por la hipótesis nula. Un p-valor pequeño indica que asumiendo que la hipótesis nula es verdadera, la diferencia entre el resultado muestral y lo predicho por la hipótesis nula no puede ser achacada al azar, por lo que deberíamos rechazar la hipótesis nula. Cuanto menor sea p-valor, menor será la probabilidad de aparición de una discrepancia como la observada, y menor credibilidad tendrá la hipótesis nula.

En la práctica, se compara el valor de p con el nivel de significación α :

- (i) Si $p \leq \alpha$ entonces rechazamos H_0 .
- (ii) Si $p > \alpha$ entonces aceptamos H_0 .
- (c) Con los p-valores comparamos áreas mientras que con el método visto con anterioridad comparamos abscisas.
- (d) En los contrastes bilaterales, el p-valor es la suma de las dos áreas, es decir, de las dos colas.

EJEMPLO 22.1 Una bióloga quiere probar, con base en la media de una muestra aleatoria de tamaño 35 y en un nivel de significación del 0.05, si la profundidad promedio del océano en cierta área es de 72.4 m., según se ha registrado. ¿Qué decidirá si obtiene $\bar{x} = 73.2$ m. y es de suponer, con base en la información recopilada en estudios similares, que $\sigma = 2.1$ m.?

Solución:

(a)

$$H_0 : \mu = 72.4$$

$$H_1 : \mu \neq 72.4$$

(b) He elegido la m.a.s. de tamaño $n=35$.

(c) Como conocemos σ , seleccionamos el estadístico de contraste

$$d = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0, 1)$$

(d) Elejimos el nivel de significación $\alpha = 0.05$.

(e) Determino la región crítica, calculando d_C tal que $p(d \geq d_C/H_0) = \alpha$, luego

$$R_C = \{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\} = \{x < -z_{0.025}\} \cup \{x > z_{0.025}\}$$

$$R_C = \{x < -1.96\} \cup \{x > 1.96\}$$

- (f) Cálculo del valor del estadístico $\hat{\theta}$ y de la medida de discrepancia d para una muestra particular, d_M ,

$$\bar{x} = 73,2; \sigma = 2,1; \mu_0 = 72,4; n = 35$$

$$d_M = \frac{73,2 - 72,4}{2,1/\sqrt{35}} = 2,25$$

- (g) Como $2,25 \in R_C$, ya que $d_M = 2,25 > d_C = 1,96$ entonces rechazamos la hipótesis nula, luego la discrepancia es estadísticamente significativa.

- (h) Obtendríamos consecuencias de tipo biológico que se quiera establecer.

Si la bióloga hubiera usado el nivel de significación del $\alpha = 0,01$ la conclusión hubiera sido diferente:

- (e) Determinamos la región crítica,

$$R_C = \{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\} = \{x < -z_{0,005}\} \cup \{x > z_{0,005}\}$$

$$R_C = \{x < -2,575\} \cup \{x > 2,575\}$$

- (f) Cálculo del valor del estadístico $\hat{\theta}$ y de la medida de discrepancia d para una muestra particular, d_M ,

$$\bar{x} = 73,2; \sigma = 2,1; \mu_0 = 72,4; n = 35$$

$$d_M = \frac{73,2 - 72,4}{2,1/\sqrt{35}} = 2,25$$

- (g) Como $2,25 \in R_A$, ya que $-2,575 < d_M = 2,25 < 2,575$ entonces aceptamos la hipótesis nula, es decir, el resultado no es significativo.

Si utilizamos el p -valor tendremos:

- (f) Cálculo del valor del estadístico $\hat{\theta}$ y de la medida de discrepancia d para una muestra particular, d_M y el p -valor para esa muestra particular

$$\bar{x} = 73,2; \sigma = 2,1; \mu_0 = 72,4; n = 35$$

$$d_M = \frac{73,2 - 72,4}{2,1/\sqrt{35}} = 2,25$$

$$p = p(d \geq d_M/H_0) = p[(d < -2,25) \cup (d > 2,25)] = 2(0,0122) = 0,0244$$

Así pues, $p=0.0244$ es el nivel de significación más bajo en que se podría rechazar la H_0 .

- (g) Podemos comparar el p -valor de este ejemplo con el nivel de significación usual

(i) Si $p = 0,0244 \leq \alpha = 0,05$ entonces rechazamos H_0 , o si,

(ii) Si $p = 0,0244 > \alpha = 0,01$ entonces aceptamos H_0 .

22.3. Relación entre intervalos y contrastes

Si por ejemplo, deseamos contrastar:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Se acepta la hipótesis nula H_0 a un nivel de significación α si es posible encontrar un intervalo $I = (U, V)$ para μ , con nivel de confianza $1 - \alpha$, que incluya a μ_0 , y viceversa.

OBSERVACIÓN 22.4

- (a) *El intervalo de confianza al nivel $1 - \alpha$ es el conjunto de hipótesis aceptables a nivel α .*
- (b) *Por tanto, $\alpha = p(\mu_0 \notin I/H_0)$.*
- (c) *El estadístico utilizado en la construcción del intervalo es la medida de discrepancia del contraste.*
- (d) *En los intervalos se contesta a la pregunta de con qué precisión conocemos el parámetro y en los contrastes a la pregunta de si el parámetro pudiera tomar un valor determinado.*
- (e) *Los intervalos tienen la ventaja frente a los contrastes de que siempre nos dan una idea de la zona en la que se va a encontrar el parámetro poblacional, mientras que en el caso de los contrastes, cuando se rechaza la hipótesis nula, no se conoce el valor del parámetro de la población en cuestión. Todo lo que se sabe es que es más verosímil que el valor del parámetro sea mayor o menor que un valor concreto.*
- (f) *Otra ventaja de los intervalos es que se pueden contrastar muchas hipótesis nulas a la vez, ya que se puede rechazar cualquier hipótesis nula de la forma $H_0 : \mu = \mu_0$ para cualquier valor de μ_0 que no pertenezca al intervalo.*

22.4. Contrastes para la media

22.4.1. Varianza conocida

Supuesto una población normal, el estadístico de discrepancia es:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0, 1)$$

	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$
Región crítica	$\{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\}$	$\{x < -z_{\alpha}\}$	$\{x > z_{\alpha}\}$

EJEMPLO 22.2 *Se quiere comprobar, con un nivel de significación del 0,05, si una muestra de tamaño 20 con $\bar{x} = 10$ procede de una población que se distribuye como una $\mathcal{N}(14, 3)$.*

Solución:

(a)

$$H_0 : \mu = 14$$

$$H_1 : \mu \neq 14$$

(b) Hemos elegido la m.a.s. de tamaño $n=20$.

(c) Como se conoce σ , selecciono el estadístico de contraste

$$d = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0, 1)$$

(d) Elejimos el nivel de significación $\alpha = 0,05$.

(e) Determinamos la región crítica, calculando d_C tal que $p(d \geq d_C/H_0) = \alpha$, luego

$$R_C = \{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\} = \{x < -z_{0,025}\} \cup \{x > z_{0,025}\}$$

$$R_C = \{x < -1,96\} \cup \{x > 1,96\}$$

(f) Cálculo del valor del estadístico $\hat{\theta}$ y de la medida de discrepancia d para una muestra particular, d_M ,

$$\bar{x} = 10, \sigma = 3, \mu_0 = 14, n = 20$$

$$d_M = \frac{10 - 14}{3/\sqrt{20}} = -5,96$$

(g) Como $-5,96 \in R_C$, ya que $d_M = -5,96 < d_C = -1,96$ entonces rechazamos la hipótesis nula, la discrepancia es significativa.

22.4.2. Varianza desconocida y muestra pequeña

Supuesto una población normal, el estadístico de discrepancia es:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n-1}} \rightsquigarrow t_{n-1}$$

	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$
R. C.	$\{x < -t_{\alpha/2;n-1}\} \cup \{x > t_{\alpha/2;n-1}\}$	$\{x < -t_{\alpha;n-1}\}$	$\{x > t_{\alpha;n-1}\}$

EJEMPLO 22.3 *Un laboratorio farmacéutico afirma que un calmante de su fabricación quita la jaqueca en como mucho 14 minutos. Con el fin de comprobar estadísticamente tal afirmación, se eligen al azar 18 pacientes con jaqueca y se toma como variable respuesta en el experimento el tiempo transcurrido desde la ingestión del calmante hasta el momento en que la jaqueca desaparece. Los resultados de la muestra han sido $\bar{x} = 19$ min. y $s = 7$ min. Se pide comprobar la afirmación del laboratorio a un nivel de significación del 0.05.*

Solución:

(a)

$$H_0 : \mu \leq 14$$

$$H_1 : \mu > 14$$

(b) Hemos elegido la m.a.s. de tamaño $n = 18 < 30$.

(c) Como desconocemos σ y tenemos una muestra pequeña, seleccionamos el estadístico de contraste

$$d = \frac{\bar{X} - \mu_0}{s/\sqrt{n-1}} \rightsquigarrow t_{n-1}$$

(d) Elejimos el nivel de significación $\alpha = 0,05$.

(e) Determinamos la región crítica, calculando d_C tal que $p(d \geq d_C/H_0) = \alpha$, luego

$$R_C = \{x > t_{\alpha;n-1}\} = \{x > t_{0,05;17}\} = \{x > 1,740\}$$

(f) Cálculo del valor del estadístico $\hat{\theta}$ y de la medida de discrepancia d para una muestra particular, d_M ,

$$\bar{x} = 19, s = 7, \mu_0 = 14, n = 18$$

$$d_M = \frac{19 - 14}{7/\sqrt{17}} = 2,94$$

(g) Como $2,94 \in R_C$, ya que $d_M = 2,94 > d_C = 1,740$ entonces rechazamos la hipótesis nula, la discrepancia es significativa.

(h) El tiempo medio de calmar la jaqueca es superior a 14 minutos con un nivel de significación del 0,05.

22.4.3. Varianza desconocida y muestra grande

Supuesto una población normal, el estadístico de discrepancia es:

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n-1}} \rightsquigarrow \mathcal{N}(0, 1)$$

	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$
R. C.	$\{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\}$	$\{x < -z_{\alpha}\}$	$\{x > z_{\alpha}\}$

EJEMPLO 22.4 *Un investigador supone que el nivel medio de una cierta sustancia en una determinada población es de 20mg./100ml. de plasma. Se toma una muestra de 40 individuos y los resultados obtenidos han sido: $\bar{x} = 18,5\text{mg./100ml.}$ y $s = 4\text{mg./100ml.}$ Se pregunta si la suposición del investigador es verdadera a un nivel de significación del 0.05.*

Solución:

(a)

$$H_0 : \mu = 20$$

$$H_1 : \mu \neq 20$$

(b) Hemos elegido la m.a.s. de tamaño $n = 40 > 30$.

(c) Como desconocemos σ y tenemos una muestra grande, seleccionamos el estadístico de contraste

$$d = \frac{\bar{X} - \mu_0}{s/\sqrt{n-1}} \rightsquigarrow \mathcal{N}(0, 1)$$

(d) Elejimos el nivel de significación $\alpha = 0,05$.

(e) Determinamos la región crítica, calculando d_C tal que $p(d \geq d_C/H_0) = \alpha$, luego

$$R_C = \{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\} = \{x < -z_{0,025}\} \cup \{x > z_{0,025}\}$$

$$R_C = \{x < -1,96\} \cup \{x > 1,96\}$$

(f) Cálculo del valor del estadístico $\hat{\theta}$ y de la medida de discrepancia d para una muestra particular, d_M ,

$$\bar{x} = 18,5; s = 4, \mu_0 = 20, n = 40$$

$$d_M = \frac{18,5 - 20}{4/\sqrt{39}} = -2,34$$

(g) Como $-2,34 \in R_C$, ya que $d_M = -2,34 < d_C = -1,96$ entonces rechazamos la hipótesis nula, la discrepancia es significativa.

(h) La suposición del investigador es falsa, con un nivel de significación del 0,05.

22.5. Contraste para la varianza

Supuesto una población normal, el estadístico de discrepancia es:

$$\chi^2 = \frac{ns^2}{\sigma_0^2} \rightsquigarrow \chi_{n-1}^2$$

	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$	$H_0 : \sigma^2 \geq \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$	$H_0 : \sigma^2 \leq \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$
R. C.	$\{x < \chi_{1-\alpha/2; n-1}^2\} \cup \{x > \chi_{\alpha/2; n-1}^2\}$	$\{x < \chi_{1-\alpha; n-1}^2\}$	$\{x > \chi_{\alpha; n-1}^2\}$

OBSERVACIÓN 22.5 *Todo contraste sobre la desviación típica se transforma en un contraste sobre la varianza, trivialmente.*

EJEMPLO 22.5 *La vida útil promedio de una muestra de 10 focos es de 4000 horas con una desviación típica de 200 horas. En general, se supone que la vida útil de los focos tiene una distribución normal. Suponga que, antes de obtener la muestra, se plantea la hipótesis de que la desviación típica $\sigma \leq 150$. Con base en los resultados muestrales, probar esa hipótesis con un nivel de significación del 1 %. Solución:*

(a)

$$H_0 : \sigma^2 \leq 150^2$$

$$H_1 : \sigma^2 > 150^2$$

(b) Hemos elegido la m.a.s. de tamaño $n = 10$.

(c) Seleccionamos el estadístico de contraste

$$d = \frac{ns^2}{\sigma_0^2} \rightsquigarrow \chi_{n-1}^2$$

(d) Elejimos el nivel de significación $\alpha = 0,01$.

(e) Determinamos la región crítica, calculando d_C tal que $p(d \geq d_C / H_0) = \alpha$, luego

$$R_C = \{x > \chi_{\alpha; n-1}^2\} = \{x > \chi_{0,01; 9}^2\} = \{x > 21,6666\}$$

(f) Cálculo del valor del estadístico $\hat{\theta}$ y de la medida de discrepancia d para una muestra particular, d_M ,

$$n = 10, s = 200, \sigma_0 = 150$$

$$d_M = \frac{ns^2}{\sigma_0^2} = \frac{10 \cdot 200^2}{150^2} = 17,7777$$

(g) Como $17,7777 \in R_A$, ya que $d_M = 17,7777 < d_C = 21,6666$ entonces aceptamos la hipótesis nula.

22.6. Contrastes para dos medias

22.6.1. Varianzas conocidas

El estadístico de discrepancia para 2 poblaciones normales independientes es:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0, 1)$$

	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
R. C.	$\{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\}$	$\{x < -z_{\alpha}\}$	$\{x > z_{\alpha}\}$

22.6.2. Varianzas desconocidas, iguales y $n_1 < 30$ ó $n_2 < 30$

El estadístico de discrepancia para 2 poblaciones normales independientes es:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 n_2}}} \rightsquigarrow t_{n_1 + n_2 - 2} = t_k$$

	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
R. C.	$\{x < -t_{\alpha/2;k}\} \cup \{x > t_{\alpha/2;k}\}$	$\{x < -t_{\alpha;k}\}$	$\{x > t_{\alpha;k}\}$

22.6.3. Varianzas desconocidas, distintas y $n_1 < 30$ ó $n_2 < 30$

El estadístico de discrepancia para 2 poblaciones normales independientes es:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}} \rightsquigarrow t_m$$

Siendo

$$c = \frac{\frac{S_1^2}{n_1 - 1}}{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}; \quad \frac{1}{m} = \frac{c^2}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1}$$

	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
R. C.	$\{x < -t_{\alpha/2;m}\} \cup \{x > t_{\alpha/2;m}\}$	$\{x < -t_{\alpha;m}\}$	$\{x > t_{\alpha;m}\}$

22.6.4. Varianzas desconocidas y $n_1 \geq 30$ ó $n_2 \geq 30$

El estadístico de discrepancia para 2 poblaciones normales independientes es:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}} \rightsquigarrow \mathcal{N}(0, 1)$$

	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
R. C.	$\{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\}$	$\{x < -z_{\alpha}\}$	$\{x > z_{\alpha}\}$

EJEMPLO 22.6 *Se sabe que la duración de una determinada enfermedad sigue una ley normal. Para la curación de dicha enfermedad se emplean dos tipos de antibióticos, el primero de ellos, elaborado por una empresa española y el segundo por una empresa norteamericana. Se observan seis enfermos a los que se le aplica el antibiótico de la empresa española y cinco enfermos a los que se les aplica el otro antibiótico. Los resultados son los siguientes:*

$$\bar{x}_1 = 12 \text{ días}, \quad s_1^2 = 16, \quad n_1 = 6$$

$$\bar{x}_2 = 15 \text{ días}, \quad s_2^2 = 16, \quad n_2 = 5$$

Se desea comprobar estadísticamente, al 0.05 de nivel de significación, si el medicamento español es igual o mejor que el norteamericano.

Solución:

(a)

$$H_0 : \mu_1 \leq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

(b) Hemos elegido las muestras de tamaño $n_1 = 6$ y $n_2 = 5$.

(c) Seleccionamos el estadístico de contraste, suponiendo la igualdad de varianzas poblacionales ya que así lo son las varianzas muestrales

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 n_2}}} \rightsquigarrow t_{n_1 + n_2 - 2} = t_k$$

(d) Elejimos el nivel de significación $\alpha = 0,05$.

(e) Determinamos la región crítica, calculando d_C tal que $p(d \geq d_C / H_0) = \alpha$, luego

$$R_C = \{x > t_{\alpha; k}\} = \{x > t_{0,05; 9}\} = \{x > 1,833\}$$

(f) Cálculo del valor de los estadísticos $\hat{\mu}_1$ y $\hat{\mu}_2$ de la medida de discrepancia para las muestras particulares, d_M ,

$$d_M = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 n_2}}} = \frac{12 - 15}{\sqrt{\frac{6(16) + 5(16)}{6 + 5 - 2} \cdot \frac{6 + 5}{6(5)}}} = -1,12$$

(g) Como $-1,12 \in R_A$, ya que $d_M = -1,12 < d_C = 1,833$ entonces aceptamos la hipótesis nula, no existe significatividad estadística.

(h) El primer antibiótico, el español, es más eficaz o tan eficaz como el segundo.

22.7. Contraste para 2 medias. Datos apareados

El estadístico de discrepancia para 2 poblaciones normales dependientes es:

$$t = \frac{\bar{d}}{S_d / \sqrt{n-1}} \rightsquigarrow t_{n-1}$$

donde

$$d_i = x_i - y_i; \quad \bar{d} = \frac{\sum d_i}{n}; \quad S_d^2 = \frac{\sum (d_i - \bar{d})^2}{n}$$

	$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 \geq \mu_2$	$H_0 : \mu_1 \leq \mu_2$
	$H_1 : \mu_1 \neq \mu_2$	$H_1 : \mu_1 < \mu_2$	$H_1 : \mu_1 > \mu_2$
R. C.	$\{x < -t_{\alpha/2; n-1}\} \cup \{x > t_{\alpha/2; n-1}\}$	$\{x < -t_{\alpha; n-1}\}$	$\{x > t_{\alpha; n-1}\}$

EJEMPLO 22.7 Una empresa farmacéutica está interesada en la investigación preliminar de un nuevo medicamento que parece tener propiedades reductoras de colesterol en sangre. A tal fin se toma una muestra al azar de 6 personas comparables, y se determina el contenido en colesterol antes y después del tratamiento. Los resultados han sido los siguientes:

Antes	217	252	229	200	209	213
Después	209	241	230	208	206	211

Se pide afirmar estadísticamente la bondad del medicamento al 0.05 de nivel de significación.

Solución:

(a)

$$H_0 : \mu_1 \leq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

(b) Hemos elegido las muestras de tamaño $n_1 = 6$ y $n_2 = 6$.

(c) Seleccionamos el estadístico de contraste,

$$t = \frac{\bar{d}}{S_d/\sqrt{n-1}} \rightsquigarrow t_{n-1}$$

(d) Elegimos el nivel de significación $\alpha = 0,05$.

(e) Determinamos la región crítica, calculando t_C tal que $p(t \geq t_C/H_0) = \alpha$, luego

$$R_C = \{x > t_{\alpha; n-1}\} = \{x > t_{0,05;5}\} = \{x > 2,015\}$$

(f) Cálculo de la medida de discrepancia para las muestras particulares,

$$t_M = \frac{\bar{d}}{s_d/\sqrt{n-1}} = \frac{2,5}{6,1305/\sqrt{5}} = 0,91$$

donde

$$\bar{d} = \frac{\sum d_i}{n} = \frac{15}{6} = 2,5$$

$$s_d^2 = \frac{\sum (d_i - \bar{d})^2}{n} = \frac{\sum d_i^2}{n} - (\bar{d})^2 = \frac{263}{6} - (2,5)^2 = 37,58 \Rightarrow s_d = 6,1305$$

(g) Como $0,91 \in R_A$, ya que $t_M = 0,91 < t_C = 2,015$ entonces aceptamos la hipótesis nula.

(h) No se puede afirmar estadísticamente la bondad del medicamento en la reducción del colesterol.

22.8. Contraste para las varianzas

El estadístico de discrepancia para 2 poblaciones normales independientes:

$$\mathcal{F} = \frac{n_1(n_2-1)S_1^2}{n_2(n_1-1)S_2^2} \rightsquigarrow \mathcal{F}_{n_1-1, n_2-1} = \mathcal{F}_{m,n}$$

	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$	$H_0 : \sigma_1^2 \geq \sigma_2^2$ $H_1 : \sigma_1^2 < \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$
R. C.	$\{x < \mathcal{F}_{1-\alpha/2; m, n}\} \cup \{x > \mathcal{F}_{\alpha/2; m, n}\}$	$\{x < \mathcal{F}_{1-\alpha; m, n}\}$	$\{x > \mathcal{F}_{\alpha; m, n}\}$

EJEMPLO 22.8 La siguiente tabla nos da el volumen en miles de Hl. exportados por dos bodegas de Jerez durante los años 2000-2004,

	2000	2001	2002	2003	2004
Sandeman	84	82	83	78	79
Croft	75	76	77	80	76

Usando el nivel de significación 0.02, ¿podemos suponer que las varianzas de las dos poblaciones de las que se efectúa el muestreo son iguales?. Y a la vista del resultado anterior, ¿cómo serían las desviaciones típicas?

Solución:

(a)

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

(b) Hemos elegido las muestras de tamaño $n_1 = n_2 = 5$.

(c) Seleccionamos el estadístico de contraste

$$d = \frac{n_1(n_2 - 1)S_1^2}{n_2(n_1 - 1)S_2^2} \rightsquigarrow \mathcal{F}_{n_1-1, n_2-1}$$

(d) Elegimos el nivel de significación $\alpha = 0,02$.

(e) Determinamos la región crítica, calculando d_C tal que $p(d \geq d_C/H_0) = \alpha$, luego

$$R_C = \{x < \mathcal{F}_{1-\alpha/2; n_1-1, n_2-1}\} \cup \{x > \mathcal{F}_{\alpha/2; n_1-1, n_2-1}\}$$

$$R_C = \{x < \mathcal{F}_{1-0,01; 4, 4}\} \cup \{x > \mathcal{F}_{0,01; 4, 4}\} = \{x < 0,06\} \cup \{x > 15,977\}$$

(f) Cálculo del valor de los estadísticos $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$ de la medida de discrepancia para las muestras particulares, d_M ,

x_i	x_i^2
84	7056
82	6724
83	6869
78	6084
79	6241
406	32994

$$\bar{x}_1 = \frac{406}{5} = 81,2; \quad \hat{\sigma}_1^2 = s_1^2 = \frac{32994}{5} - (81,2)^2 = 5,36$$

x_i	x_i^2
75	5625
76	5776
77	5929
80	6400
76	5776
384	29506

$$\bar{x}_2 = \frac{384}{5} = 76,8; \quad \hat{\sigma}_2^2 = s_2^2 = \frac{29506}{5} - (76,8)^2 = 2,96$$

$$d_M = \frac{n_1(n_2 - 1)s_1^2}{n_2(n_1 - 1)s_2^2} = \frac{5(5 - 1)5,36}{5(5 - 1)2,96} = 1,8108$$

(g) Como $1,8108 \in R_A$, ya que $0,06 < d_M = 1,8108 < d_C = 15,977$ entonces aceptamos la hipótesis nula, esto es, con un nivel de significación del 0,02 existe evidencia de que las varianzas son iguales y por tanto de igualdad de desviaciones típicas.

22.9. Contrastes asintóticos

Los siguientes contrastes son asintóticos en el sentido de que la distribución del estadístico la obtendremos pasando al límite y por tanto necesitaremos tamaños de muestras suficientemente grandes.

22.9.1. Contraste para la proporción

El estadístico de discrepancia es:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}} \rightsquigarrow \mathcal{N}(0, 1) \quad \text{si } n \rightarrow \infty$$

	$H_0 : p = p_0$ $H_1 : p \neq p_0$	$H_0 : p \geq p_0$ $H_1 : p < p_0$	$H_0 : p \leq p_0$ $H_1 : p > p_0$
R. C.	$\{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\}$	$\{x < -z_{\alpha}\}$	$\{x > z_{\alpha}\}$

EJEMPLO 22.9 *Un crítico de televisión sostiene que el 80 % de todos los televidentes considera que el nivel de ruido de cierto anuncio es demasiado alto. Si una muestra aleatoria de 320 televidentes incluye 245 que encuentran que el ruido del anuncio es demasiado alto, pruebe en el nivel 0.05 de significación si la proporción muestral confirma lo predicho por el crítico de televisión.*

Solución:

(a)

$$H_0 : p = 0,8$$

$$H_1 : p \neq 0,8$$

(b) Hemos elegido la m.a.s. de tamaño $n = 320 > 30$.

(c) Seleccionamos el estadístico de contraste asintótico

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}} \rightsquigarrow \mathcal{N}(0, 1) \quad \text{si } n \rightarrow \infty$$

(d) Elejimos el nivel de significación $\alpha = 0,05$.(e) Determinamos la región crítica, calculando d_C tal que $p(d \geq d_C/H_0) = \alpha$, luego

$$R_C = \{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\} = \{x < -z_{0,025}\} \cup \{x > z_{0,025}\}$$

$$R_C = \{x < -1,96\} \cup \{x > 1,96\}$$

(f) Cálculo del valor del estadístico $\hat{p}$ y de la medida de discrepancia d para una muestra particular, d_M ,

$$\hat{p} = \frac{245}{320} = 0,7656; p_0 = 0,8; n = 320$$

$$d_M = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}} = \frac{0,7656 - 0,8}{\sqrt{\frac{0,7656(1 - 0,7656)}{320}}} = -1,44$$

(g) Como $-1,44 \in R_A$, ya que $-1,96 = d_C < d_M = -1,44 < d_C = 1,96$ entonces aceptamos la hipótesis nula, no existe significatividad estadística.

(h) La proporción muestral confirma lo predicho por el crítico, con un nivel de significación del 0,05.

22.9.2. Contraste para la igualdad de proporciones

El estadístico de discrepancia es:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}} \rightsquigarrow \mathcal{N}(0, 1) \quad \text{si } n_1 \rightarrow \infty, \quad n_2 \rightarrow \infty$$

	$H_0 : p_1 = p_2$ $H_1 : p_1 \neq p_2$	$H_0 : p_1 \geq p_2$ $H_1 : p_1 < p_2$	$H_0 : p_1 \leq p_2$ $H_1 : p_1 > p_2$
R. C.	$\{x < -z_{\alpha/2}\} \cup \{x > z_{\alpha/2}\}$	$\{x < -z_{\alpha}\}$	$\{x > z_{\alpha}\}$

EJEMPLO 22.10 Se quiere comprobar la efectividad de una determinada vacuna contra la alergia. Para ello se suministró la vacuna a 100 pacientes y se les comparó con un grupo testigo de 100 pacientes también afectados por la alergia en épocas pasadas, no vacunadas.

Entre los vacunados, 8 sufrieron alergias y entre los no vacunados, 25 volvieron a sufrir alergias. ¿Podemos afirmar al 0.05 de nivel de significación que la vacuna es eficaz contra la alergia?

Solución:

(a)

$$H_0 : p_1 \geq p_2$$

$$H_1 : p_1 < p_2$$

(b) Hemos elegido las muestras de tamaño $n_1 = n_2 = 100 > 30$.

(c) Seleccionamos el estadístico de contraste, que de una manera asintótica tiende a una normal

$$d = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}} \rightsquigarrow \mathcal{N}(0, 1)$$

(d) Elejimos el nivel de significación $\alpha = 0,05$.

(e) Determinamos la región crítica, calculando d_C tal que $p(d \geq d_C/H_0) = \alpha$, luego

$$R_C = \{x < -z_{\alpha}\} = \{x < -1,65\}$$

(f) Cálculo del valor de los estadísticos $\hat{p}_1$ y $\hat{p}_2$ de la medida de discrepancia para las muestras particulares, d_M ,

$$d_M = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}} = \frac{0,08 - 0,25}{\sqrt{\frac{0,08(1 - 0,08)}{100} + \frac{0,25(1 - 0,25)}{100}}} = -3,3$$

(g) Como $-3,3 \in R_C$, ya que $d_M = -3,3 < -1,65 = d_C$ entonces rechazamos la hipótesis nula.

(h) La vacuna es eficaz.

Cuestiones, problemas y recursos

Cuestiones

1. Define población y muestra. Ejemplos.
2. ¿Qué son muestreos probabilísticos? Cita sus ventajas e inconvenientes.
3. ¿Qué es la Inferencia Estadística? Cita sus técnicas más importantes.
4. Cita las ventajas e inconvenientes de la realización de una encuesta frente a un censo.
5. Explica brevemente en qué consiste el muestreo por conglomerados con submuestreo.
6. Ventajas e inconvenientes del muestreo estratificado.
7. Ventajas e inconvenientes del muestreo por conglomerados.
8. Diferencias entre conglomerados y estratos.
9. La siguiente tabla recoge las notas de 30 estudiantes en la asignatura Estadística Administrativa, siendo los 15 primeros chicos y los 15 segundos chicas y los 5 primeros de Sanlúcar y los restantes de Jerez:

1	4	3	2	8	2	3	9	1	4	1	3	6	9	8
4	6	8	8	9	7	2	2	3	9	4	7	1	3	9

Estima la proporción de estudiantes que aprueban, si dicha estimación se basa:

- (a) En una muestra aleatoria de tamaño 5.
- (b) En una muestra estratificada de tamaño 6 de afijación uniforme.
- (c) En una muestra sistemática de tamaño 5.

- (d) En una muestra por conglomerados de tamaño 1.
10. Los siguientes datos corresponden con salarios diarios en euros de una población de 50 personas, siendo las 10 primeras mujeres y los restantes hombres; y los 10 primeros de Sanlúcar, los 10 siguientes de Puerto Real, los siguientes 10 de Trebujena, los 10 siguientes de Sevilla y los restantes de Jerez:

64	76	75	84	78	79	79	62	70	73
75	78	75	66	80	73	86	78	87	80
75	60	99	91	74	79	67	78	83	85
86	84	84	70	73	78	76	81	76	82
78	68	97	62	86	77	73	60	65	80

Estima el salario medio en toda la población, si dicha estimación se basa:

- En una muestra aleatoria de tamaño 10.
 - En una muestra estratificada de tamaño 10 de afijación proporcional.
 - En una muestra sistemática de tamaño 10.
 - En una muestra por conglomerados de tamaño 2.
 - En una muestra por conglomerados bietápico de tamaño 2 en primera etapa y 3 en segunda.
11. ¿Qué es la afijación? Tipos.
12. Diferencias y semejanzas entre la t de Student y la Normal.
13. ¿En qué consiste la estimación puntual?
14. ¿Qué es un estadístico? Ejemplos.
15. Define el nivel de significación en los intervalos.
16. ¿Qué interpretación tiene el nivel de confianza?
17. ¿Es verdad la afirmación de que a mayor nivel de confianza mayor precisión?
18. Cita formas de mejorar la precisión en los intervalos de confianza.
19. ¿Qué son los intervalos asintóticos? Cita alguno que lo sea.
20. ¿En qué consiste el método analógico?
21. ¿Qué es el p -valor?
22. Da un ejemplo de un test unilateral.
23. Diferencias entre intervalos de confianza y contrastes.

Problemas

1. Realizado un estudio para probar si existe alguna diferencia entre alturas promedio de las niñas nacidas en dos países distintos, Irlanda y España, muestras aleatorias dieron los siguientes resultados:

$$n_I = 12, \quad \bar{x}_I = 62,7, \quad s_I = 2,5$$

$$n_E = 15, \quad \bar{x}_E = 61,8, \quad s_E = 2,62$$

- (a) Construye un intervalo de confianza al 95 % para la altura media de la población de jóvenes de España.
 - (b) Determinar un intervalo de confianza para la diferencia de medias manteniendo el nivel de confianza, y si se conoce por estudios anteriores que las dos varianzas poblacionales son iguales.
 - (c) A la vista del anterior intervalo, ¿podríamos afirmar que las estaturas de las dos poblaciones de jóvenes son iguales?
2. Se seleccionan dos muestras aleatorias e independientes que nos dan el número de puestos de trabajo creados en el último mes por diferentes empresas de dos sectores económicos. Los datos de las muestras son los siguientes:

$$n_A = 6, \quad \bar{x}_A = 16,1666, \quad s_A = 8,1410$$

$$n_B = 6, \quad \bar{x}_B = 22,6666, \quad s_B = 23,5733$$

Con el fin de conocer el impacto de las nuevas modalidades de contratación en ambos sectores y suponiendo que el número de empleos creados en ambos sectores siguiera sendas distribuciones normales con varianzas iguales:

- (a) Determina un intervalo de confianza para la desviación típica poblacional del número de empleos del sector B, a un nivel de confianza del 95 %.
 - (b) ¿Se puede admitir, al nivel de confianza anterior, que la desviación típica del número de empleos del sector B es igual a 2? Razónalo.
 - (c) Determina un intervalo de confianza para la diferencia del número medio de empleos entre las dos poblaciones correspondientes a los dos sectores, a un nivel de confianza del 99 %.
 - (d) A la vista del anterior intervalo, ¿podríamos afirmar que ambos sectores son similares en cuanto al número medio de empleos creados en el último mes, al nivel de confianza anterior?
3. Dos técnicas de alumbrado distintas se comparan midiendo la intensidad de luz en áreas seleccionadas. Si 11 lecturas de la primera técnica tienen una desviación típica de 2.6 y 16 lecturas de la segunda técnica tienen una desviación típica de 4.4.

- (a) Determina un intervalo de confianza para la desviación típica poblacional de la primera técnica, a un nivel de confianza del 95 %.
 - (b) Determinar un intervalo de confianza al 95 % para el cociente de varianzas. A la vista del intervalo anterior, ¿son igualmente variables las dos técnicas? Razónalo.
4. El diario "EL INDISCRETO" hace una encuesta para comparar los salarios semanales, en euros, de varias profesiones en dos zonas de España. Para ello se han emparejado las diferentes profesiones por regiones como aquí se indica:

	A	B	C	D	E	F	G
Andalucía	228	203	197	283	283	267	358
Cataluña	267	207	209	291	324	265	382

Con la información anterior se requiere:

- (a) Determina un intervalo de confianza para la diferencia de salarios semanales medios a un nivel de confianza del 95 %.
 - (b) A la vista del intervalo anterior, ¿los salarios de las dos regiones se pueden considerar iguales? Si no es así, dí en qué región el salario es mayor. Razona las respuestas.
 - (c) ¿Cuál es el error máximo que se comete para el salario semanal medio de Andalucía al 95 % de confianza y qué sugieres para reducirlo?
5. Los datos del siguiente cuadro reflejan las pérdidas semanales de horas/hombre como consecuencia de accidentes en 10 obras de Jerez antes y después de instaurar cierto programa de seguridad:

Antes	45	73	46	124	33	57	83	34	26	17
Después	36	60	44	119	35	51	77	29	24	11

- (a) Usa el nivel de significación del 0.05 para construir un intervalo de confianza para la diferencia de pérdidas semanales de horas/hombre antes y después del programa.
 - (b) A la vista del intervalo anterior, ¿es eficaz el programa de seguridad? Razona la respuesta.
6. El Ayuntamiento de Jerez decide impartir unos cursos para mejorar la atención al público de los comerciantes del centro. Seleccionados al azar 8 comercios, se recogen en la tabla adjunta las calificaciones otorgadas por la inspección y obtenidas por dichos comercios antes y después de los cursos:

Comercio	A	B	C	D	E	F	G	H
Antes	3.2	3.3	3.4	2.1	4.1	3.1	2.9	4.2
Después	3.1	3.5	3.6	3	4.2	3.3	2.5	4

Con la información anterior se requiere:

- (a) Calcular el intervalo de confianza para la nota media antes de la realización del curso al 95 % de nivel de confianza.
 - (b) Determina el intervalo de confianza al 90 % para la diferencia de notas medias. ¿Podríamos extraer alguna conclusión definitiva sobre si la decisión de hacer los cursos mejora las puntuaciones otorgadas por la inspección?
7. El Vicerrector de la Universidad de Cádiz recopila datos de una m.a.s. de 230 estudiantes inscritos en los programas de postgrado de Administración de Empresas, y encuentra que 54 de ellos tienen la licenciatura en Economía. Estime la proporción de esos estudiantes a nivel de toda la universidad que son licenciados en Economía, utilizando un intervalo de confianza del 90 %.
8. La empresa de investigación de mercados RODRIGUEZ Y COTILLAS ASOCIADOS, entrevistan a 100 lectores de periódicos, vecinos de Jerez, y encuentran una proporción del 40 % entre los que preferían el DIARIO DE JEREZ a otros diarios. Con la información anterior se requiere:
- (a) Determina mediante un intervalo de confianza del 95 % la proporción de lectores de todo Jerez que prefieren el DIARIO DE JEREZ.
 - (b) Supóngase que antes de recopilar los datos se especificó que la estimación por intervalo del 95 % de confianza debería estar dentro del $\pm 0,05$ y se dio como estimación previa de p un valor igual a 0.3. ¿Cuál debería ser el tamaño de la muestra?
 - (c) Contesta a la anterior pregunta si se desconoce cualquier posible estimación de p . ¿Por qué aumenta el tamaño de la muestra?
9. El diario EL PAÍS realiza una encuesta a 1500 votantes de Andalucía, de los que 752 manifestaron su apoyo al PSOE, y a 1000 votantes de Extremadura, de los que 548 se inclinan por el PSOE.
- (a) Calcula un intervalo de confianza al 95 % para la diferencia de proporciones de votantes del PSOE en las dos comunidades autónomas.
 - (b) ¿Pueden considerar los dirigentes del PSOE que la proporción de votantes que expresan su intención de votarles es igual en las dos comunidades autónomas?

- (c) Si nos fijamos sólo en Andalucía, cuántas entrevistas debería realizar el periódico para cometer un error en la proporción de votos que recibiría el PSOE de un 2 % como máximo, manteniendo el nivel de confianza y bajo el supuesto de la máxima indeterminación de la proporción.
10. La producción de lino de 6 parcelas de la comarca de Jerez es de 1.4, 1.8, 1.1, 1.9, 2.2 y 1.2 toneladas por hectárea, respectivamente. Pruebe al nivel de significación del 5 % si esto respalda el argumento del Comisario de Agricultura de Bruselas de que la producción promedio para el lino es de 1.5 toneladas por hectárea, primero mediante la técnica de los contrastes y después mediante la de los intervalos de confianza.
11. La cadena de pizzerías TELEPIZZA pretende abrir una nueva pizzería en la calle Guarnidos de Jerez si al menos 200 automóviles pasan por el lugar propuesto cada hora, por término medio, durante determinadas horas. Para 20 horas muestreadas al azar, dentro del intervalo especificado, se encuentra que el número promedio de automóviles que pasan por ese lugar es de $\bar{x} = 208$ con $s = 30$. Se supone que la población estadística es aproximadamente una normal.
- (a) ¿Abrirán una nueva pizzería si el nivel de significación es del 5 %?
- (b) Contestar a la anterior pregunta utilizando la técnica del p -valor.
12. Para una muestra aleatoria de $n_1 = 10$ focos, se encuentra que la vida promedio es de $\bar{x}_1 = 4000$ horas con $s_1 = 200$. Para otra marca de focos, para los cuales se supone también una vida útil con distribución normal y varianza distinta a la anterior, una muestra de $n_2 = 8$ focos tiene una media muestral de $\bar{x}_2 = 4300$ y una desviación típica de $s_2 = 250$. Contrastar la hipótesis de que no existe diferencia significativa entre la vida útil promedio de las dos marcas de focos, utilizando un nivel de significación del 1 %.
13. Una muestra de 100 bombillas de la marca A dan una vida media de 1190 horas y desviación típica 90 horas. Una muestra de 75 bombillas de la marca B dan una vida media de 1230 horas y desviación típica de 120 horas. ¿Hay diferencia entre las vidas medias de estas dos marcas de bombillas al nivel de significación del 5 % primero y del 1 % después?
14. Las siguientes muestras aleatorias son las ganancias diarias por trabajador de cinco categorías distintas, de Jerez y Arcos. Las ganancias siguen sendas normales con varianzas iguales:

Jerez	8400	8230	8380	7860	7930
Arcos	7510	7690	7720	8070	7660

- (a) Use el nivel de significación del 5 % para probar si las medias de ambas poblaciones son iguales.

(b) Idem utilizando la técnica del p -valor.

15. El fabricante de automóviles OPEL obtiene datos de rendimiento de gasolina para una muestra de 9 automóviles en diversas categorías de peso utilizando gasolina común, con y sin aditivo. Por supuesto se utilizan los mismos conductores para las dos condiciones (de hecho, el conductor no sabe qué tipo de gasolina se utiliza en las pruebas). Con los datos del rendimiento de la tabla siguiente, probar la hipótesis de que no existe diferencia significativa entre el kilometraje promedio que se obtiene con y sin aditivo, utilizando el nivel de significación del 5 %.

Coche	1	2	3	4	5	6	7	8	9
Km. CA	9.41	9.18	8.18	7.51	7.28	6.59	6.21	5.79	5.62
Km. SA	9.28	9.15	8.28	7.59	7.21	6.62	6.13	5.64	5.51

16. De la Facultad de Ciencias Sociales y de la Comunicación de Jerez de la Frontera, una muestra aleatoria de 11 estudiantes tiene un promedio de calificación de 2.7 con una desviación típica de 0.4. Si cogemos 10 estudiantes de la Facultad de Ciencias del Trabajo, éstos tienen una calificación media de 2.9 y una desviación típica de 0.3. Se supone que las calificaciones tienen una distribución normal.

- (a) Usa el nivel de significación del 10 % para probar la hipótesis de que las dos varianzas poblacionales son iguales.
- (b) Prueba la hipótesis de que la calificación promedio para las dos Facultades es distinta para el anterior nivel de significación.

17. Se plantea la hipótesis de que como mucho el 5 % de los relojes PELUCO tienen defectos. Para una muestra de 100 de tales relojes se encuentran que 10 están defectuosos.

- (a) Prueba la hipótesis nula al 5 % de nivel de significación.
- (b) Idem utilizando la técnica del p -valor.

18. Una muestra de 50 hogares de la zona de FEDERICO MAYO arroja que 10 de ellos se encuentran viendo el programa "Cifras y Letras", en ese momento. En la zona de LOS NARANJOS, 15 hogares de los 50 seleccionados están viendo el citado programa. Se quiere probar la hipótesis de que la proporción de telespectadores en las dos zonas es igual utilizando el nivel de significación del 1 %.

19. Para probar la efectividad de un nuevo analgésico, se suministró a 80 pacientes de una clínica una píldora que contenía analgésico y a otros 80 pacientes se les dio un placebo que sólo contenía azúcar. Si 56 pacientes del primer grupo y 38 del segundo grupo sintieron efecto benéfico:

- (a) ¿Qué podemos concluir al nivel de significación del 1 % acerca de la efectividad del medicamento?
 - (b) ¿Utilizando la técnica del p -valor la conclusión es similar?
20. Estudios realizados entre alumnos de último curso del colegio Lasalle en 1960 y de nuevo en 2000, demostraron que en 1960 la altura promedio de 400 chicos era de 168 cm. con una desviación típica de 2,6, mientras que en 2000 la altura promedio de 400 chicos del mismo grupo de edad era de 170 cm. con una desviación típica de 2.5.
- (a) Calcula un intervalo de confianza al 90 % para la diferencia de alturas medias entre los dos períodos de tiempo.
 - (b) A la vista del anterior intervalo, ¿podemos considerar las dos medias poblacionales iguales?
 - (c) Usa la técnica de los contrastes de significación para contrastar al nivel de significación del 0,10 si las dos alturas medias poblacionales de ambos períodos de tiempo coinciden.
21. Se toma una muestra aleatoria de 16 bonos de un tipo y la varianza de los vencimientos fue de 123,35 y para otra muestra independiente de 11 bonos de un segundo tipo, la varianza de los vencimientos fue de 8,02. Asumiendo que las dos poblaciones tienen distribución normal.
- (a) Calcula un intervalo de confianza al 95 % para la razón de varianzas poblacionales.
 - (b) A raíz del anterior intervalo, ¿es posible aceptar que las dos varianzas poblacionales son iguales?.
 - (c) Utilizando la técnica de los contrastes de significación, contrasta la hipótesis de la igualdad de varianzas al 5 % de significación.
22. Para verificar la aseveración del servicio 061 de que por lo menos el 40 % de todas sus llamadas son emergencias de vida o muerte, se tomó una muestra aleatoria de todos sus registros y se encontró que 49 de 150 llamadas fueron de emergencias de vida o muerte.
- (a) Contrasta la aseveración del servicio 061 al 5 % de significación.
 - (b) Calcula un intervalo de confianza del 95 % para la proporción de llamadas que recibe el servicio 061.
 - (c) Calcula el tamaño muestral necesario para cometer un error máximo de 0,02 al 95 % de confianza tomando como $\hat{p} = 0,3$.

Recursos

Seguidamente exponemos una serie de páginas de la web donde podemos encontrar algunos recursos electrónicos relacionados con esta parte:

- (a) [http : //www.ruf.rice.edu/ ~ lane/stat_sim/sampling_dist/index.html](http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html) simula la extracción de muestras de una población normal y va construyendo histogramas con los valores de las medias de las muestras.
- (b) [http : //www.grad.cgs.edu/wise/sdmmod/sdm.html](http://www.grad.cgs.edu/wise/sdmmod/sdm.html) muestra la distribución de la media muestral para ciertos tipos de población. Simula la extracción de muestras y dibuja histogramas que permiten comparar como se parece la distribución obtenida por simulación con la prevista teóricamente.
- (c) [http : //www.arbitrage.byu.edu/sample.html](http://www.arbitrage.byu.edu/sample.html) de la Brigham Young University, se compara como evoluciona el histograma y la media de una muestra a medida que se le van añadiendo nuevas observaciones. Se compara la media muestral con la media poblacional, para constatar que esas diferencias tienden a reducirse cuando aumenta el tamaño de la muestra.
- (d) [http : //www.stat.sc.edu/~ west/javahtml/ConfidenceInterval.html](http://www.stat.sc.edu/~west/javahtml/ConfidenceInterval.html) se calculan intervalos de confianza para medir a partir de muestras generadas de una $\mathcal{N}(0, 1)$. Permite cambiar de manera fácil el nivel de confianza y cuentan con intervalos que no incluyen el verdadero valor del parámetro.
- (e) [http : //www.grad.cgs.edu/wise/hypothesis/hypoth_applet.html](http://www.grad.cgs.edu/wise/hypothesis/hypoth_applet.html) dibuja la población de la que se extrae la hipótesis nula y otra de la que se extraerán las muestras. Simula y representa la extracción de muestras y presenta los resultados del contraste de hipótesis.

Bibliografía

- [1] Alba Fernández, V. y Muñoz Vázquez, A. (2000). *Introducción a la Estadística Pública*, Jaén: Universidad de Jaén. Servicio de publicaciones.
- [2] Arenales Abad, C. y Rodríguez Ruiz, J. (1988). *Problemas de Estadísticas Económicas*, Madrid: Editorial Pirámide, S.A.
- [3] Bedate, A., González, J., Rivas, A. y Sanz, J.A., (1996). *Problemas de Estadística Descriptiva Empresarial*, Barcelona: Editorial Ariel.
- [4] Behar Gutiérrez, R. y Grima Cintas, P. (2000). Selección de recursos en Internet para la Enseñanza-Aprendizaje de la Estadística, *Boletín de la SEIO*, **16 (4)**, 24-28.
- [5] Berenson, M.L. y otros. (2001). *Estadística para la Administración*, Madrid: Prentice-Hall.
- [6] Calvo Gómez, F. y Sarramona López, J. (1983). *Ejercicios de Estadística Aplicada a las Ciencias Sociales*, Madrid: Editorial CEAC.
- [7] Casas Sánchez, J.M. y Santos Peña, J. (2002). *Introducción a la Estadística para la Economía y Administración de Empresas*, Madrid: Editorial Centro de Estudios Ramón Areces, S.A.
- [8] Coquillat Durán, F. (1991). *Estadística Descriptiva. Metodología y Cálculo*, Madrid: Editorial Tebar Flores.
- [9] Durá Peiró, J.M. y López Cuñat, J.M. (1992). *Fundamentos de la Estadística. Estadística Descriptiva y Modelos Probabilísticos por la Inferencia*, Barcelona: Editorial Ariel.
- [10] Fernández Abascal, H. y otros. (1994). *Cálculo de Probabilidades y Estadística*, Barcelona: Editorial Ariel, S.A.
- [11] Fernández Cuesta, C. y Fuentes García, F. (1995). *Curso de Estadística Descriptiva. Teoría y Práctica*, Barcelona: Editorial Ariel, S.A.

- [12] García Ramos, J.A., Ramos González, C.D. y Ruiz Garzón, G. (1998). *Estadística Aplicada I*, Cádiz: Copistería San Rafael, S.L.
- [13] Jurado Bello, M. y Diz Pérez, J. (2002). Una revisión de recursos de Internet para la docencia de Estadística, *Boletín de la SEIO*, **18 (3 y 4)**, 2-8.
- [14] Labrousse, C. (1973). *Estadística. Ejercicios resueltos. Tomo I*, Madrid: Editorial Paraninfo.
- [15] Labrousse, C. (1973). *Estadística. Ejercicios resueltos. Tomo II*, Madrid: Editorial Paraninfo.
- [16] Levin, R. (1988). *Estadística para Administradores*, Madrid: Prentice-Hall.
- [17] Martín Pliego, F.J. (1994). *Introducción a la Estadística Económica y Empresarial (Teoría y práctica)*, Madrid: Editorial AC.
- [18] Martín Pliego, F.J. y Ruiz-Maya Pérez, L. (1997). *Estadística. I: Probabilidad*, Madrid: Editorial AC.
- [19] Martín Pliego, F.J. y Ruiz-Maya Pérez, L. (1997). *Estadística. II: Inferencia*, Madrid: Editorial AC.
- [20] Montiel Torres, A.M., Rius Díaz, F. y Barón López, F.J. (1998). *Elementos Básicos de Estadística Económica y Empresarial*, Madrid: Prentice Hall.
- [21] Peña, D. (2001). *Fundamentos de Estadística*, Madrid: Alianza Editorial.
- [22] Pérez C. (1995). *Curso Estadístico con Statgraphics. Técnicas básicas*, Madrid: Editorial RA-MA.
- [23] Quesada Paloma, V., Isidoro Martín, A. y López Martín, L.A. (1990). *Curso y Ejercicios de Estadística*, Madrid: Editorial Alhambra Longman, S.A.
- [24] Ríos, S. (1989). *Ejercicios de Estadística*, Madrid: Editorial Paraninfo.
- [25] Ruiz-Maya, L. (1986). *Problemas de Estadística*, Madrid: Editorial AC.
- [26] Spiegel, M.R. (1997). *Estadística*, Madrid: Editorial McGraw-Hill.



MANUALES MATEMÁTICAS Y FÍSICA

Presentamos este manual fruto de nuestra labor pedagógica en la Diplomatura de Gestión y Administración y Pública, y con el deseo fundamental de que resulte útil a, no sólo a nuestros estudiantes, sino a cualquier estudiante de Primer Ciclo universitario, donde un curso básico de Estadística tenga cabida. Este volumen que el lector tiene en sus manos, viene a cubrir la práctica inexistencia de manuales dedicados a la Estadística Administrativa como tal, debido entre otros factores, a la singularidad que la Diplomatura de Gestión y Administración Pública tiene en la Universidad Española.

No obstante, la relación entre la Administración y la Estadística ha sido muy intensa. No olvidemos que los términos Estadística y Estado comparten la misma raíz latina o que la Estadística fue llamada en sus comienzos Política Aritmética.

Esta obra se ha estructurado en seis partes, finalizando cada una de ellas con una colección de problemas y cuestiones propuestas, que proporcionan la oportunidad de repetir los principios básicos de la asignatura, elemento fundamental para un aprendizaje eficaz de la misma, así como algunas direcciones electrónicas dónde encontrar interesantes recursos.

Las primeras partes de este ejemplar están dedicadas a la descripción de variables con modelos frecuentistas, los números índices y las series temporales, conformando lo que se entiende por Estadística Descriptiva.

La quinta parte la dedicaremos a la construcción de modelos de distribuciones de probabilidad que pudieran representar el comportamiento de diferentes fenómenos aleatorios que aparecen en el mundo real.

Dedicaremos la última parte al estudio de la Inferencia Estadística, comprendiendo técnicas como la estimación y el contraste de hipótesis, con objeto de proporcionar al investigador instrumentos para la toma de decisiones cuando prevalece el azar. También dedicaremos un capítulo a la introducción al muestreo en poblaciones finitas y otro al estudio de las características de las principales encuestas que realiza la Administración.



UCA

Universidad
de Cádiz

Servicio de Publicaciones



9 788498 280661